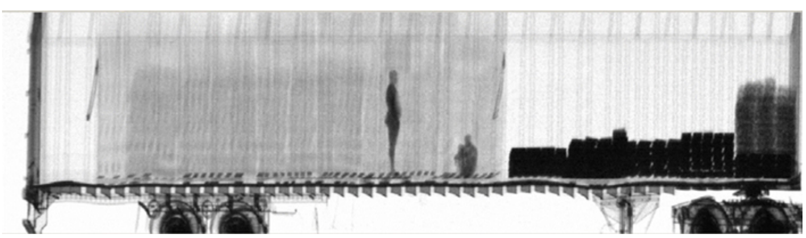
Introduction to Applications of Nuclear Physics class="introduction"

- Provide examples of various nuclear physics applications.

Tori Randall, Ph.D., curator for the Department of Physical Anthropology at the San Diego Museum of Man, prepares a 550-year-old Peruvian child mummy for a CT scan at Naval Medical Center San Diego. (credit: U.S. Navy photo by Mass Communication Specialist 3rd Class Samantha A. Lewis)

Applications of nuclear physics have become an integral part of modern life. From the bone scan that detects a cancer to the radioiodine treatment that cures another, nuclear radiation has diagnostic and therapeutic effects on medicine. From the fission power reactor to the hope of controlled fusion, nuclear energy is now commonplace and is a part of our plans for the future. Yet, the destructive potential of nuclear weapons haunts us, as does the possibility of nuclear reactor accidents. Certainly, several applications of nuclear physics escape our view, as seen in [link]. Not only has nuclear physics revealed secrets of nature, it has an inevitable impact based on its applications, as they are intertwined with human values. Because of its potential for alleviation of suffering, and its power as an ultimate destructor of life, nuclear physics is often viewed with ambivalence. But it provides perhaps the best example that applications can be good or evil, while knowledge itself is neither.

Customs officers inspect vehicles using neutron irradiation. Cars and trucks pass through portable x-ray machines that reveal their contents. (credit: Gerald L. Nino, CBP, U.S. Dept. of Homeland Security)

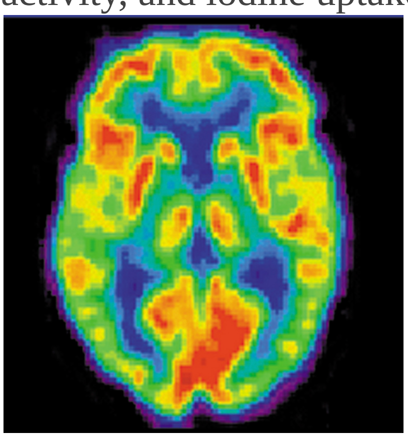


This image shows two stowaways caught illegally entering the United States from Canada. (credit: U.S. Customs and Border Protection)

Medical Imaging and Diagnostics

- Explain the working principle behind an anger camera.
- Describe the SPECT and PET imaging techniques.

A host of medical imaging techniques employ nuclear radiation. What makes nuclear radiation so useful? First, γ radiation can easily penetrate tissue; hence, it is a useful probe to monitor conditions inside the body. Second, nuclear radiation depends on the nuclide and not on the chemical compound it is in, so that a radioactive nuclide can be put into a compound designed for specific purposes. The compound is said to be **tagged**. A tagged compound used for medical purposes is called a **radiopharmaceutical**. Radiation detectors external to the body can determine the location and concentration of a radiopharmaceutical to yield medically useful information. For example, certain drugs are concentrated in inflamed regions of the body, and this information can aid diagnosis and treatment as seen in [link]. Another application utilizes a radiopharmaceutical which the body sends to bone cells, particularly those that are most active, to detect cancerous tumors or healing points. Images can then be produced of such bone scans. Radioisotopes are also used to determine the functioning of body organs, such as blood flow, heart muscle activity, and iodine uptake in the thyroid gland.



A radiopharmaceutical is used to produce this brain image of a patient with

Alzheimer's disease. Certain features are computer enhanced. (credit: National Institutes of Health)

Medical Application

[link] lists certain medical diagnostic uses of radiopharmaceuticals, including isotopes and activities that are typically administered. Many organs can be imaged with a variety of nuclear isotopes replacing a stable element by a radioactive isotope. One common diagnostic employs iodine to image the thyroid, since iodine is concentrated in that organ. The most active thyroid cells, including cancerous cells, concentrate the most iodine and, therefore, emit the most radiation. Conversely, hypothyroidism is indicated by lack of iodine uptake. Note that there is more than one isotope that can be used for several types of scans. Another common nuclear diagnostic is the thallium scan for the cardiovascular system, particularly used to evaluate blockages in the coronary arteries and examine heart activity. The salt TlCl can be used, because it acts like NaCl and follows the blood. Gallium-67 accumulates where there is rapid cell growth, such as in tumors and sites of infection. Hence, it is useful in cancer imaging. Usually, the patient receives the injection one day and has a whole body scan 3 or 4 days later because it can take several days for the gallium to build up.

Procedure, isotope	**Typical activity (mCi), where** $1\ \text{mCi} = 3.7 \times 10^{7}\ \text{Bq}$

Procedure, isotope	Typical activity (mCi), where $1 \text{ mCi} = 3.7 \times 10^7 \text{ Bq}$
Brain scan	
^{99m}Tc	7.5
^{113m}In	7.5
$^{11}\text{C (PET)}$	20
$^{13}\text{N (PET)}$	20
$^{15}\text{O (PET)}$	50
$^{18}\text{F (PET)}$	10
Lung scan	
^{99m}Tc	2

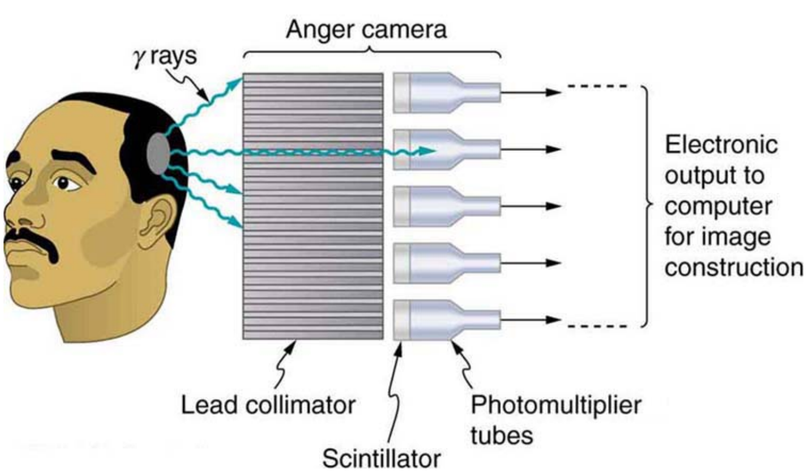
Procedure, isotope	Typical activity (mCi), where $1 \text{ mCi} = 3.7 \times 10^7 \text{ Bq}$
^{133}Xe	7.5
Cardiovascular blood pool	
^{131}I	0.2
$^{99\text{m}}\text{Tc}$	2
Cardiovascular arterial flow	
^{201}Tl	3
^{24}Na	7.5
Thyroid scan	
^{131}I	0.05
^{123}I	0.07

Procedure, isotope	**Typical activity (mCi), where** $1\ \mathrm{mCi} = 3.7 \times 10^7\ \mathrm{Bq}$
Liver scan	
$^{198}\mathrm{Au}$ (colloid)	0.1
$^{99\mathrm{m}}\mathrm{Tc}$ (colloid)	2
Bone scan	
$^{85}\mathrm{Sr}$	0.1
$^{99\mathrm{m}}\mathrm{Tc}$	10
Kidney scan	
$^{197}\mathrm{Hg}$	0.1
$^{99\mathrm{m}}\mathrm{Tc}$	1.5

Diagnostic Uses of Radiopharmaceuticals

Note that [link] lists many diagnostic uses for ${}^{99m}\mathrm{Tc}$, where “m” stands for a metastable state of the technetium nucleus. Perhaps 80 percent of all radiopharmaceutical procedures employ ${}^{99m}\mathrm{Tc}$ because of its many advantages. One is that the decay of its metastable state produces a single, easily identified 0.142-MeV γ ray. Additionally, the radiation dose to the patient is limited by the short 6.0-h half-life of ${}^{99m}\mathrm{Tc}$. And, although its half-life is short, it is easily and continuously produced on site. The basic process for production is neutron activation of molybdenum, which quickly β decays into ${}^{99m}\mathrm{Tc}$. Technetium-99m can be attached to many compounds to allow the imaging of the skeleton, heart, lungs, kidneys, etc.

[link] shows one of the simpler methods of imaging the concentration of nuclear activity, employing a device called an **Anger camera** or **gamma camera**. A piece of lead with holes bored through it collimates γ rays emerging from the patient, allowing detectors to receive γ rays from specific directions only. The computer analysis of detector signals produces an image. One of the disadvantages of this detection method is that there is no depth information (i.e., it provides a two-dimensional view of the tumor as opposed to a three-dimensional view), because radiation from any location under that detector produces a signal.



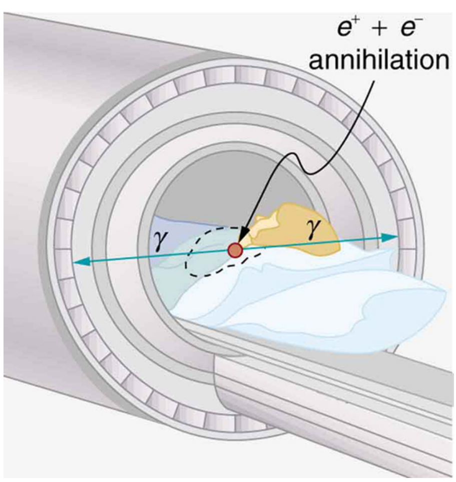
An Anger or gamma camera consists of a lead collimator and an array of detectors. Gamma rays produce light flashes in the

scintillators. The light output is converted to an electrical signal by the photomultipliers. A computer constructs an image from the detector output.

Imaging techniques much like those in x-ray computed tomography (CT) scans use nuclear activity in patients to form three-dimensional images. [link] shows a patient in a circular array of detectors that may be stationary or rotated, with detector output used by a computer to construct a detailed image. This technique is called **single-photon-emission computed tomography(SPECT)** or sometimes simply SPET. The spatial resolution of this technique is poor, about 1 cm, but the contrast (i.e. the difference in visual properties that makes an object distinguishable from other objects and the background) is good.

SPECT uses a geometry similar to a CT scanner to form an image of the concentration of a radiopharmaceutical compound. (credit: Woldo, Wikimedia Commons)

Images produced by β^+ emitters have become important in recent years. When the emitted positron (β^+) encounters an electron, mutual annihilation occurs, producing two γ rays. These γ rays have identical 0.511-MeV energies (the energy comes from the destruction of an electron or positron mass) and they move directly away from one another, allowing detectors to determine their point of origin accurately, as shown in [link]. The system is called **positron emission tomography (PET)**. It requires detectors on opposite sides to simultaneously (i.e., at the same time) detect photons of 0.511-MeV energy and utilizes computer imaging techniques similar to those in SPECT and CT scans. Examples of β^+ -emitting isotopes used in PET are ^{11}C, ^{13}N, ^{15}O, and ^{18}F, as seen in [link]. This list includes C, N, and O, and so they have the advantage of being able to function as tags for natural body compounds. Its resolution of 0.5 cm is better than that of SPECT; the accuracy and sensitivity of PET scans make them useful for examining the brain's anatomy and function. The brain's use of oxygen and water can be monitored with ^{15}O. PET is used extensively for diagnosing brain disorders. It can note decreased metabolism in certain regions prior to a confirmation of Alzheimer's disease. PET can locate regions in the brain that become active when a person carries out specific activities, such as speaking, closing their eyes, and so on.



A PET system takes

advantage of the two identical γ-ray photons produced by positron-electron annihilation. These γ rays are emitted in opposite directions, so that the line along which each pair is emitted is determined. Various events detected by several pairs of detectors are then analyzed by the computer to form an accurate image.

Note:
PhET Explorations: Simplified MRI

Is it a tumor? Magnetic Resonance Imaging (MRI) can tell. Your head is full of tiny radio transmitters (the nuclear spins of the hydrogen nuclei of your water molecules). In an MRI unit, these little radios can be made to broadcast their positions, giving a detailed picture of the inside of your head.

Simplified MRI

Section Summary

- Radiopharmaceuticals are compounds that are used for medical imaging and therapeutics.
- The process of attaching a radioactive substance is called tagging.
- [link] lists certain diagnostic uses of radiopharmaceuticals including the isotope and activity typically used in diagnostics.
- One common imaging device is the Anger camera, which consists of a lead collimator, radiation detectors, and an analysis computer.
- Tomography performed with γ-emitting radiopharmaceuticals is called SPECT and has the advantages of x-ray CT scans coupled with organ- and function-specific drugs.
- PET is a similar technique that uses β^+ emitters and detects the two annihilation γ rays, which aid to localize the source.

Conceptual Questions

Exercise:

Problem:

In terms of radiation dose, what is the major difference between medical diagnostic uses of radiation and medical therapeutic uses?

Exercise:

Problem:

One of the methods used to limit radiation dose to the patient in medical imaging is to employ isotopes with short half-lives. How would this limit the dose?

Problems & Exercises

Exercise:

Problem:

A neutron generator uses an α source, such as radium, to bombard beryllium, inducing the reaction ${}^{4}\text{He} + {}^{9}\text{Be} \rightarrow {}^{12}\text{C} + n$. Such neutron sources are called RaBe sources, or PuBe sources if they use plutonium to get the α s. Calculate the energy output of the reaction in MeV.

Solution:

5.701 MeV

Exercise:

Problem:

Neutrons from a source (perhaps the one discussed in the preceding problem) bombard natural molybdenum, which is 24 percent ${}^{98}\text{Mo}$. What is the energy output of the reaction ${}^{98}\text{Mo} + n \rightarrow {}^{99}\text{Mo} + \gamma$? The mass of ${}^{98}\text{Mo}$ is given in Appendix A: Atomic Masses, and that of ${}^{99}\text{Mo}$ is 98.907711 u.

Exercise:

Problem:

The purpose of producing ${}^{99}\text{Mo}$ (usually by neutron activation of natural molybdenum, as in the preceding problem) is to produce ${}^{99\text{m}}\text{Tc}$. Using the rules, verify that the β^{-} decay of ${}^{99}\text{Mo}$ produces ${}^{99\text{m}}\text{Tc}$. (Most ${}^{99\text{m}}\text{Tc}$ nuclei produced in this decay are left in a metastable excited state denoted ${}^{99\text{m}}\text{Tc}$.)

Solution:

${}^{99}_{42}\text{Mo}_{57} \rightarrow {}^{99}_{43}\text{Tc}_{56} + \beta^{-} + v_e$

Exercise:

Problem:

(a) Two annihilation γ rays in a PET scan originate at the same point and travel to detectors on either side of the patient. If the point of origin is 9.00 cm closer to one of the detectors, what is the difference in arrival times of the photons? (This could be used to give position information, but the time difference is small enough to make it difficult.)

(b) How accurately would you need to be able to measure arrival time differences to get a position resolution of 1.00 mm?

Exercise:

Problem:

[link] indicates that 7.50 mCi of ^{99m}Tc is used in a brain scan. What is the mass of technetium?

Solution:

1.43×10^{-9} g

Exercise:

Problem:

The activities of ^{131}I and ^{123}I used in thyroid scans are given in [link] to be 50 and 70 μCi, respectively. Find and compare the masses of ^{131}I and ^{123}I in such scans, given their respective half-lives are 8.04 d and 13.2 h. The masses are so small that the radioiodine is usually mixed with stable iodine as a carrier to ensure normal chemistry and distribution in the body.

Exercise:

Problem:

(a) Neutron activation of sodium, which is 100% ^{23}Na, produces ^{24}Na, which is used in some heart scans, as seen in [link]. The equation for the reaction is $^{23}\text{Na} + n \rightarrow {}^{24}\text{Na} + \gamma$. Find its energy output, given the mass of ^{24}Na is 23.990962 u.

(b) What mass of ^{24}Na produces the needed 5.0-mCi activity, given its half-life is 15.0 h?

Solution:

(a) 6.958 MeV

(b) 5.7×10^{-10} g

Glossary

Anger camera
: a common medical imaging device that uses a scintillator connected to a series of photomultipliers

gamma camera
: another name for an Anger camera

positron emission tomography (PET)
: tomography technique that uses β^{+} emitters and detects the two annihilation γ rays, aiding in source localization

radiopharmaceutical
: compound used for medical imaging

single-photon-emission computed tomography (SPECT)
: tomography performed with γ-emitting radiopharmaceuticals

tagged
: process of attaching a radioactive substance to a chemical compound

Biological Effects of Ionizing Radiation

- Define various units of radiation.
- Describe RBE.

We hear many seemingly contradictory things about the biological effects of ionizing radiation. It can cause cancer, burns, and hair loss, yet it is used to treat and even cure cancer. How do we understand these effects? Once again, there is an underlying simplicity in nature, even in complicated biological organisms. All the effects of ionizing radiation on biological tissue can be understood by knowing that **ionizing radiation affects molecules within cells, particularly DNA molecules.**

Let us take a brief look at molecules within cells and how cells operate. Cells have long, double-helical DNA molecules containing chemical codes called genetic codes that govern the function and processes undertaken by the cell. It is for unraveling the double-helical structure of DNA that James Watson, Francis Crick, and Maurice Wilkins received the Nobel Prize. Damage to DNA consists of breaks in chemical bonds or other changes in the structural features of the DNA chain, leading to changes in the genetic code. In human cells, we can have as many as a million individual instances of damage to DNA per cell per day. It is remarkable that DNA contains codes that check whether the DNA is damaged or can repair itself. It is like an auto check and repair mechanism. This repair ability of DNA is vital for maintaining the integrity of the genetic code and for the normal functioning of the entire organism. It should be constantly active and needs to respond rapidly. The rate of DNA repair depends on various factors such as the cell type and age of the cell. A cell with a damaged ability to repair DNA, which could have been induced by ionizing radiation, can do one of the following:

- The cell can go into an irreversible state of dormancy, known as senescence.
- The cell can commit suicide, known as programmed cell death.
- The cell can go into unregulated cell division leading to tumors and cancers.

Since ionizing radiation damages the DNA, which is critical in cell reproduction, it has its greatest effect on cells that rapidly reproduce, including most types of cancer. Thus, cancer cells are more sensitive to radiation than normal cells and can be killed by it easily. Cancer is characterized by a malfunction of cell reproduction, and can also be caused by ionizing radiation. Without contradiction, ionizing radiation can be both a cure and a cause.

To discuss quantitatively the biological effects of ionizing radiation, we need a radiation dose unit that is directly related to those effects. All effects of radiation are assumed to be directly proportional to the amount of ionization produced in the biological organism. The amount of ionization is in turn proportional to the amount of deposited energy. Therefore, we define a **radiation dose unit** called the **rad**, as $1/100$ of a joule of ionizing energy deposited per kilogram of tissue, which is

Equation:

$$1 \text{ rad} = 0.01 \text{ J/kg}.$$

For example, if a 50.0-kg person is exposed to ionizing radiation over her entire body and she absorbs 1.00 J, then her whole-body radiation dose is

Equation:

$$(1.00 \text{ J})/(50.0 \text{ kg}) = 0.0200 \text{ J/kg} = 2.00 \text{ rad}.$$

If the same 1.00 J of ionizing energy were absorbed in her 2.00-kg forearm alone, then the dose to the forearm would be

Equation:

$$(1.00 \text{ J})/(2.00 \text{ kg}) = 0.500 \text{ J/kg} = 50.0 \text{ rad},$$

and the unaffected tissue would have a zero rad dose. While calculating radiation doses, you divide the energy absorbed by the mass of affected tissue. You must specify the affected region, such as the whole body or forearm in addition to giving the numerical dose in rads. The SI unit for radiation dose is the **gray (Gy)**, which is defined to be

Equation:

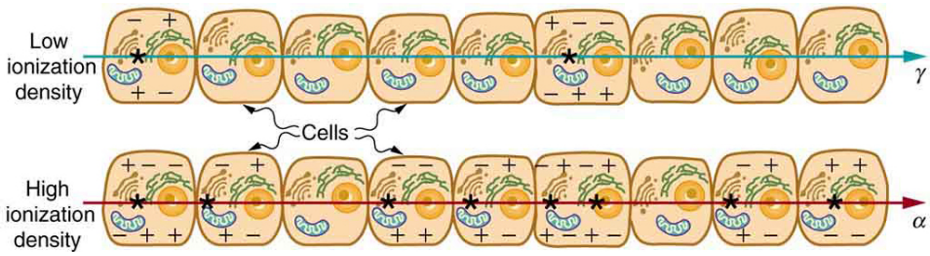
$$1 \text{ Gy} = 1 \text{ J/kg} = 100 \text{ rad}.$$

However, the rad is still commonly used. Although the energy per kilogram in 1 rad is small, it has significant effects since the energy causes ionization. The energy needed for a single ionization is a few eV, or less than 10^{-18} J. Thus, 0.01 J of ionizing energy can create a huge number of ion pairs and have an effect at the cellular level.

The effects of ionizing radiation may be directly proportional to the dose in rads, but they also depend on the type of radiation and the type of tissue. That is, for a given dose in rads, the effects depend on whether the radiation is α, β, γ, x-ray, or some other type of ionizing radiation. In the earlier discussion of the range of ionizing radiation, it was noted that energy is deposited in a series of ionizations and not in a single interaction. Each ion pair or ionization requires a certain amount of energy, so that the number of ion pairs is directly proportional to the amount of the deposited ionizing energy. But, if the range of the radiation is small, as it is for α s, then the ionization and the damage created is more concentrated and harder for the organism to repair, as seen in [link]. Concentrated damage is more difficult for biological organisms to repair than damage that is spread out, so short-range particles have greater biological effects. The **relative biological effectiveness** (RBE) or **quality factor** (QF) is given in [link] for several types of ionizing radiation—the effect of the radiation is directly proportional to the RBE. A dose unit more closely related to effects in biological tissue is called the **roentgen equivalent man** or rem and is defined to be the dose in rads multiplied by the relative biological effectiveness.

Equation:

$$\text{rem} = \text{rad} \times \text{RBE}$$



The image shows ionization created in cells by α and γ radiation. Because of its shorter range, the ionization and damage created by α is more concentrated and harder for the organism to repair. Thus, the RBE for α s is greater than the RBE for γ s, even though they create the same amount of ionization at the same energy.

So, if a person had a whole-body dose of 2.00 rad of γ radiation, the dose in rem would be $(2.00 \text{ rad})(1) = 2.00 \text{ rem whole body}$. If the person had a whole-body dose of 2.00 rad of α radiation, then the dose in rem would be $(2.00 \text{ rad})(20) = 40.0 \text{ rem whole body}$. The α s would have 20 times the effect on the person than the γ s for the same deposited energy. The SI equivalent of the rem is the **sievert** (Sv), defined to be $\text{Sv} = \text{Gy} \times \text{RBE}$, so that

Equation:

$$1 \text{ Sv} = 1 \text{ Gy} \times \text{RBE} = 100 \text{ rem}.$$

The RBEs given in [link] are approximate, but they yield certain insights. For example, the eyes are more sensitive to radiation, because the cells of the lens do not repair themselves. Neutrons cause more damage than γ rays, although both are neutral and have large ranges, because neutrons often cause secondary radiation when they are captured. Note that the RBEs are 1 for higher-energy β s, γ s, and x-rays, three of the most common types of radiation. For those types of radiation, the numerical values of the dose in rem and rad are identical. For example, 1 rad of γ radiation is also 1 rem. For that reason, rads are still widely quoted rather than rem. [link] summarizes the units that are used for radiation.

Note:

Misconception Alert: Activity vs. Dose
“Activity” refers to the radioactive source while “dose” refers to the amount of energy from the radiation that is deposited in a person or object.

A high level of activity doesn’t mean much if a person is far away from the source. The activity R of a source depends upon the quantity of material (kg) as well as the half-life. A short half-life will produce many more disintegrations per second. Recall that $R = \frac{0.693N}{t_{1/2}}$. Also, the activity decreases exponentially, which is seen in the equation $R = R_0 e^{-\lambda t}$.

Type and energy of radiation	**RBE[footnote]** **Values approximate, difficult to determine.**
X-rays	1
γ rays	1
β rays greater than 32 keV	1
βrays less than 32 keV	1.7
Neutrons, thermal to slow (<20 keV)	2–5
Neutrons, fast (1–10 MeV)	10 (body), 32 (eyes)
Protons (1–10 MeV)	10 (body), 32 (eyes)
α rays from radioactive decay	10–20
Heavy ions from accelerators	10–20

Relative Biological Effectiveness

Quantity	SI unit name	Definition	Former unit	Conversion
Activity	Becquerel (bq)	decay/sec	Curie (Ci)	$1\text{ Bq} = 2.7 \times 10^{-11}\text{ Ci}$
Absorbed dose	Gray (Gy)	1 J/kg	rad	$\text{Gy} = 100\text{ rad}$
Dose Equivalent	Sievert (Sv)	1 J/kg × RBE	rem	$\text{Sv} = 100\text{ rem}$

Units for Radiation

The large-scale effects of radiation on humans can be divided into two categories: immediate effects and long-term effects. [link] gives the immediate effects of whole-body exposures received in less than one day. If the radiation exposure is spread out over more time, greater doses are needed to cause the effects listed. This is due to the body's ability to partially repair the damage. Any dose less than 100 mSv (10 rem) is called a **low dose**, 0.1 Sv to 1 Sv (10 to 100 rem) is called a **moderate dose**, and anything greater than 1 Sv (100 rem) is called a **high dose**. There is no known way to determine after the fact if a person has been exposed to less than 10 mSv.

Dose in Sv [footnote] Multiply by 100 to obtain dose in rem.	**Effect**
0–0.10	No observable effect.
0.1 – 1	Slight to moderate decrease in white blood cell counts.
0.5	Temporary sterility; 0.35 for women, 0.50 for men.
1 – 2	Significant reduction in blood cell counts, brief nausea and vomiting. Rarely fatal.
2 – 5	Nausea, vomiting, hair loss, severe blood damage, hemorrhage, fatalities.

Dose in Sv [footnote] Multiply by 100 to obtain dose in rem.	Effect
4.5	LD50/32. Lethal to 50% of the population within 32 days after exposure if not treated.
5 – 20	Worst effects due to malfunction of small intestine and blood systems. Limited survival.
>20	Fatal within hours due to collapse of central nervous system.

Immediate Effects of Radiation (Adults, Whole Body, Single Exposure)

Immediate effects are explained by the effects of radiation on cells and the sensitivity of rapidly reproducing cells to radiation. The first clue that a person has been exposed to radiation is a change in blood count, which is not surprising since blood cells are the most rapidly reproducing cells in the body. At higher doses, nausea and hair loss are observed, which may be due to interference with cell reproduction. Cells in the lining of the digestive system also rapidly reproduce, and their destruction causes nausea. When the growth of hair cells slows, the hair follicles become thin and break off. High doses cause significant cell death in all systems, but the lowest doses that cause fatalities do so by weakening the immune system through the loss of white blood cells.

The two known long-term effects of radiation are cancer and genetic defects. Both are directly attributable to the interference of radiation with cell reproduction. For high doses of radiation, the risk of cancer is reasonably well known from studies of exposed groups. Hiroshima and Nagasaki survivors and a smaller number of people exposed by their occupation, such as radium dial painters, have been fully documented. Chernobyl victims will be studied for many decades, with some data already available. For example, a significant increase in childhood thyroid cancer has been observed. The risk of a radiation-induced cancer for low and moderate doses is generally *assumed* to be proportional to the risk known for high doses. Under this assumption, any dose of radiation, no matter how small, involves a risk to human health. This is called the **linear hypothesis** and it may be prudent, but it *is* controversial. There is some evidence that, unlike the immediate effects of radiation, the long-term effects are cumulative and there is little self-repair. This is analogous to the risk of skin cancer from UV exposure, which is known to be cumulative.

There is a latency period for the onset of radiation-induced cancer of about 2 years for leukemia and 15 years for most other forms. The person is at risk for at least 30 years after the latency period. Omitting many details, the overall risk of a radiation-induced cancer

death per year per rem of exposure is about 10 in a million, which can be written as $10/10^6\ \text{rem}\cdot\text{y}$.

If a person receives a dose of 1 rem, his risk each year of dying from radiation-induced cancer is 10 in a million and that risk continues for about 30 years. The lifetime risk is thus 300 in a million, or 0.03 percent. Since about 20 percent of all worldwide deaths are from cancer, the increase due to a 1 rem exposure is impossible to detect demographically. But 100 rem (1 Sv), which was the dose received by the average Hiroshima and Nagasaki survivor, causes a 3 percent risk, which can be observed in the presence of a 20 percent normal or natural incidence rate.

The incidence of genetic defects induced by radiation is about one-third that of cancer deaths, but is much more poorly known. The lifetime risk of a genetic defect due to a 1 rem exposure is about 100 in a million or $3.3/10^6\ \text{rem}\cdot\text{y}$, but the normal incidence is 60,000 in a million. Evidence of such a small increase, tragic as it is, is nearly impossible to obtain. For example, there is no evidence of increased genetic defects among the offspring of Hiroshima and Nagasaki survivors. Animal studies do not seem to correlate well with effects on humans and are not very helpful. For both cancer and genetic defects, the approach to safety has been to use the linear hypothesis, which is likely to be an overestimate of the risks of low doses. Certain researchers even claim that low doses are *beneficial*. **Hormesis** is a term used to describe generally favorable biological responses to low exposures of toxins or radiation. Such low levels may help certain repair mechanisms to develop or enable cells to adapt to the effects of the low exposures. Positive effects may occur at low doses that could be a problem at high doses.

Even the linear hypothesis estimates of the risks are relatively small, and the average person is not exposed to large amounts of radiation. [link] lists average annual background radiation doses from natural and artificial sources for Australia, the United States, Germany, and world-wide averages. Cosmic rays are partially shielded by the atmosphere, and the dose depends upon altitude and latitude, but the average is about 0.40 mSv/y. A good example of the variation of cosmic radiation dose with altitude comes from the airline industry. Monitored personnel show an average of 2 mSv/y. A 12-hour flight might give you an exposure of 0.02 to 0.03 mSv.

Doses from the Earth itself are mainly due to the isotopes of uranium, thorium, and potassium, and vary greatly by location. Some places have great natural concentrations of uranium and thorium, yielding doses ten times as high as the average value. Internal doses come from foods and liquids that we ingest. Fertilizers containing phosphates have potassium and uranium. So we are all a little radioactive. Carbon-14 has about 66 Bq/kg radioactivity whereas fertilizers may have more than 3000 Bq/kg radioactivity. Medical and dental diagnostic exposures are mostly from x-rays. It should be noted that x-ray doses tend to be localized and are becoming much smaller with improved techniques. [link] shows typical doses received during various diagnostic x-ray examinations. Note the large dose from a CT scan. While CT scans only account for less than 20 percent of

the x-ray procedures done today, they account for about 50 percent of the annual dose received.

Radon is usually more pronounced underground and in buildings with low air exchange with the outside world. Almost all soil contains some ^{226}Ra and ^{222}Rn, but radon is lower in mainly sedimentary soils and higher in granite soils. Thus, the exposure to the public can vary greatly, even within short distances. Radon can diffuse from the soil into homes, especially basements. The estimated exposure for ^{222}Rn is controversial. Recent studies indicate there is more radon in homes than had been realized, and it is speculated that radon may be responsible for 20 percent of lung cancers, being particularly hazardous to those who also smoke. Many countries have introduced limits on allowable radon concentrations in indoor air, often requiring the measurement of radon concentrations in a house prior to its sale. Ironically, it could be argued that the higher levels of radon exposure and their geographic variability, taken with the lack of demographic evidence of any effects, means that low-level radiation is *less* dangerous than previously thought.

Radiation Protection

Laws regulate radiation doses to which people can be exposed. The greatest occupational whole-body dose that is allowed depends upon the country and is about 20 to 50 mSv/y and is rarely reached by medical and nuclear power workers. Higher doses are allowed for the hands. Much lower doses are permitted for the reproductive organs and the fetuses of pregnant women. Inadvertent doses to the public are limited to $1/10$ of occupational doses, except for those caused by nuclear power, which cannot legally expose the public to more than $1/1000$ of the occupational limit or 0.05 mSv/y (5 mrem/y). This has been exceeded in the United States only at the time of the Three Mile Island (TMI) accident in 1979. Chernobyl is another story. Extensive monitoring with a variety of radiation detectors is performed to assure radiation safety. Increased ventilation in uranium mines has lowered the dose there to about 1 mSv/y.

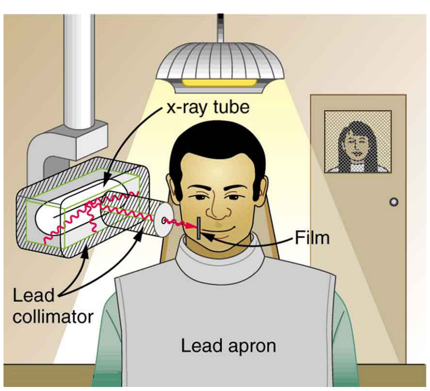
Source	**Dose (mSv/y)[footnote]** **Multiply by 100 to obtain dose in mrem/y.**			
Source	Australia	Germany	United States	World
Natural Radiation - external				

Source	Dose (mSv/y)[footnote] Multiply by 100 to obtain dose in mrem/y.			
Cosmic Rays	0.30	0.28	0.30	0.39
Soil, building materials	0.40	0.40	0.30	0.48
Radon gas	0.90	1.1	2.0	1.2
Natural Radiation - internal				
^{40}K, ^{14}C, ^{226}Ra	0.24	0.28	0.40	0.29
Medical & Dental	0.80	0.90	0.53	0.40
TOTAL	2.6	3.0	3.5	2.8

Background Radiation Sources and Average Doses

To physically limit radiation doses, we use **shielding**, increase the **distance** from a source, and limit the **time of exposure**.

[link] illustrates how these are used to protect both the patient and the dental technician when an x-ray is taken. Shielding absorbs radiation and can be provided by any material, including sufficient air. The greater the distance from the source, the more the radiation spreads out. The less time a person is exposed to a given source, the smaller is the dose received by the person. Doses from most medical diagnostics have decreased in recent years due to faster films that require less exposure time.



A lead apron is placed over the dental patient and shielding surrounds the x-ray tube to limit exposure to tissue other than the tissue that is being imaged. Fast films limit the time needed to obtain images, reducing exposure to the imaged tissue. The technician stands a few meters away behind a lead-lined door with a lead glass window, reducing her occupational exposure.

Procedure	**Effective dose (mSv)**
Chest	0.02
Dental	0.01
Skull	0.07
Leg	0.02
Mammogram	0.40
Barium enema	7.0
Upper GI	3.0
CT head	2.0
CT abdomen	10.0

Typical Doses Received During Diagnostic X-ray Exams

Problem-Solving Strategy

You need to follow certain steps for dose calculations, which are

Step 1. *Examine the situation to determine that a person is exposed to ionizing radiation.*

Step 2. *Identify exactly what needs to be determined in the problem (identify the unknowns).* The most straightforward problems ask for a dose calculation.

Step 3. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).* Look for information on the type of radiation, the energy per event, the activity, and the mass of tissue affected.

Step 4. *For dose calculations, you need to determine the energy deposited.* This may take one or more steps, depending on the given information.

Step 5. *Divide the deposited energy by the mass of the affected tissue.* Use units of joules for energy and kilograms for mass. If a dose in Sv is involved, use the definition that $1\ \text{Sv} = 1\ \text{J/kg}$.

Step 6. *If a dose in mSv is involved, determine the RBE (QF) of the radiation.* Recall that $1\ \text{mSv} = 1\ \text{mGy} \times \text{RBE}$ (or $1\ \text{rem} = 1\ \text{rad} \times \text{RBE}$).

Step 7. *Check the answer to see if it is reasonable: Does it make sense?* The dose should be consistent with the numbers given in the text for diagnostic, occupational, and therapeutic exposures.

Example:
Dose from Inhaled Plutonium
Calculate the dose in rem/y for the lungs of a weapons plant employee who inhales and retains an activity of $1.00\ \mu\text{Ci}$ of ^{239}Pu in an accident. The mass of affected lung tissue is 2.00 kg, the plutonium decays by emission of a 5.23-MeV α particle, and you may assume the higher value of the RBE for α s from [link].
Strategy
Dose in rem is defined by $1\ \text{rad} = 0.01\ \text{J/kg}$ and $\text{rem} = \text{rad} \times \text{RBE}$. The energy deposited is divided by the mass of tissue affected and then multiplied by the RBE. The latter two quantities are given, and so the main task in this example will be to find the energy deposited in one year. Since the activity of the source is given, we can calculate the number of decays, multiply by the energy per decay, and convert MeV to joules to get the total energy.
Solution

The activity $R = 1.00\ \mu\text{Ci} = 3.70 \times 10^4\ \text{Bq} = 3.70 \times 10^4$ decays/s. So, the number of decays per year is obtained by multiplying by the number of seconds in a year:
Equation:

$$\left(3.70 \times 10^4\ \text{decays/s}\right)\left(3.16 \times 10^7\ \text{s}\right) = 1.17 \times 10^{12}\ \text{decays.}$$

Thus, the ionizing energy deposited per year is
Equation:

$$E = \left(1.17 \times 10^{12}\ \text{decays}\right)\left(5.23\ \text{MeV/decay}\right) \times \left(\frac{1.60 \times 10^{-13}\ \text{J}}{\text{MeV}}\right) = 0.978\ \text{J.}$$

Dividing by the mass of the affected tissue gives
Equation:

$$\frac{E}{\text{mass}} = \frac{0.978\ \text{J}}{2.00\ \text{kg}} = 0.489\ \text{J/kg.}$$

One Gray is 1.00 J/kg, and so the dose in Gy is
Equation:

$$\text{dose in Gy} = \frac{0.489\ \text{J/kg}}{1.00\ (\text{J/kg})/\text{Gy}} = 0.489\ \text{Gy.}$$

Now, the dose in Sv is
Equation:

$$\text{dose in Sv} = \text{Gy} \times \text{RBE}$$

Equation:

$$= (0.489\ \text{Gy})(20) = 9.8\ \text{Sv.}$$

Discussion
First note that the dose is given to two digits, because the RBE is (at best) known only to two digits. By any standard, this yearly radiation dose is high and will have a devastating effect on the health of the worker. Worse yet, plutonium has a long radioactive half-life and is not readily eliminated by the body, and so it will remain in the lungs. Being an α emitter makes the effects 10 to 20 times worse than the same ionization produced by β s, γ rays, or x-rays. An activity of 1.00 μCi is created by only 16 μg of ^{239}Pu (left as an end-of-chapter problem to verify), partly justifying claims that plutonium is the most toxic substance known. Its actual hazard depends on how likely it is to be spread out among a large population and then ingested. The Chernobyl disaster's deadly legacy, for example, has nothing to do with the plutonium it put into the environment.

Risk versus Benefit

Medical doses of radiation are also limited. Diagnostic doses are generally low and have further lowered with improved techniques and faster films. With the possible exception of routine dental x-rays, radiation is used diagnostically only when needed so that the low risk is justified by the benefit of the diagnosis. Chest x-rays give the lowest doses—about 0.1 mSv to the tissue affected, with less than 5 percent scattering into tissues that are not directly imaged. Other x-ray procedures range upward to about 10 mSv in a CT scan, and about 5 mSv (0.5 rem) per dental x-ray, again both only affecting the tissue imaged. Medical images with radiopharmaceuticals give doses ranging from 1 to 5 mSv, usually localized. One exception is the thyroid scan using ^{131}I. Because of its relatively long half-life, it exposes the thyroid to about 0.75 Sv. The isotope ^{123}I is more difficult to produce, but its short half-life limits thyroid exposure to about 15 mSv.

Note:
PhET Explorations: Alpha Decay
Watch alpha particles escape from a polonium nucleus, causing radioactive alpha decay. See how random decay times relate to the half life.

Alpha Decay

Section Summary

- The biological effects of ionizing radiation are due to two effects it has on cells: interference with cell reproduction, and destruction of cell function.
- A radiation dose unit called the rad is defined in terms of the ionizing energy deposited per kilogram of tissue:
  Equation:
  $$1\ \text{rad} = 0.01\ \text{J/kg}.$$
- The SI unit for radiation dose is the gray (Gy), which is defined to be $1\ \text{Gy} = 1\ \text{J/kg} = 100\ \text{rad}.$
- To account for the effect of the type of particle creating the ionization, we use the relative biological effectiveness (RBE) or quality factor (QF) given in [link] and define a unit called the roentgen equivalent man (rem) as

Equation:

$$\mathrm{rem} = \mathrm{rad} \times \mathrm{RBE}.$$

- Particles that have short ranges or create large ionization densities have RBEs greater than unity. The SI equivalent of the rem is the sievert (Sv), defined to be
Equation:

$$\mathrm{Sv} = \mathrm{Gy} \times \mathrm{RBE} \text{ and } 1 \text{ Sv} = 100 \text{ rem}.$$

- Whole-body, single-exposure doses of 0.1 Sv or less are low doses while those of 0.1 to 1 Sv are moderate, and those over 1 Sv are high doses. Some immediate radiation effects are given in [link]. Effects due to low doses are not observed, but their risk is assumed to be directly proportional to those of high doses, an assumption known as the linear hypothesis. Long-term effects are cancer deaths at the rate of $10/10^6 \text{ rem}\cdot\text{y}$and genetic defects at roughly one-third this rate. Background radiation doses and sources are given in [link]. World-wide average radiation exposure from natural sources, including radon, is about 3 mSv, or 300 mrem. Radiation protection utilizes shielding, distance, and time to limit exposure.

Conceptual Questions

Exercise:

Problem:

Isotopes that emit α radiation are relatively safe outside the body and exceptionally hazardous inside. Yet those that emit γ radiation are hazardous outside and inside. Explain why.

Exercise:

Problem:

Why is radon more closely associated with inducing lung cancer than other types of cancer?

Exercise:

Problem:

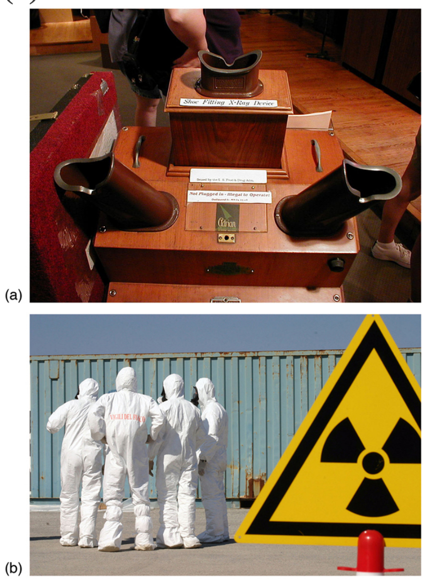
The RBE for low-energy βs is 1.7, whereas that for higher-energy βs is only 1. Explain why, considering how the range of radiation depends on its energy.

Exercise:

Problem:

Which methods of radiation protection were used in the device shown in the first photo in [link]? Which were used in the situation shown in the second photo?

(a)



(a) This x-ray fluorescence machine is one of the thousands used in shoe stores to produce images of feet as a check on the fit of shoes. They are unshielded and remain on as long as the feet are in them, producing doses much greater than medical images. Children were fascinated with them. These machines were used in shoe stores until laws preventing such unwarranted radiation exposure were enacted in the 1950s. (credit: Andrew Kuchling) (b) Now that we know the effects of exposure to

radioactive material, safety is a priority. (credit: U.S. Navy)

Exercise:

Problem:

What radioisotope could be a problem in homes built of cinder blocks made from uranium mine tailings? (This is true of homes and schools in certain regions near uranium mines.)

Exercise:

Problem:

Are some types of cancer more sensitive to radiation than others? If so, what makes them more sensitive?

Exercise:

Problem:

Suppose a person swallows some radioactive material by accident. What information is needed to be able to assess possible damage?

Problems & Exercises

Exercise:

Problem:

What is the dose in mSv for: (a) a 0.1 Gy x-ray? (b) 2.5 mGy of neutron exposure to the eye? (c) 1.5 mGy of α exposure?

Solution:

(a) 100 mSv

(b) 80 mSv

(c) ~30 mSv

Exercise:

Problem:

Find the radiation dose in Gy for: (a) A 10-mSv fluoroscopic x-ray series. (b) 50 mSv of skin exposure by an α emitter. (c) 160 mSv of β^- and γ rays from the ^{40}K in your body.

Exercise:

Problem:

How many Gy of exposure is needed to give a cancerous tumor a dose of 40 Sv if it is exposed to α activity?

Solution:

~2 Gy

Exercise:

Problem:

What is the dose in Sv in a cancer treatment that exposes the patient to 200 Gy of γ rays?

Exercise:

Problem:

One half the γ rays from ^{99m}Tc are absorbed by a 0.170-mm-thick lead shielding. Half of the γ rays that pass through the first layer of lead are absorbed in a second layer of equal thickness. What thickness of lead will absorb all but one in 1000 of these γ rays?

Solution:

1.69 mm

Exercise:

Problem:

A plumber at a nuclear power plant receives a whole-body dose of 30 mSv in 15 minutes while repairing a crucial valve. Find the radiation-induced yearly risk of death from cancer and the chance of genetic defect from this maximum allowable exposure.

Exercise:

Problem:

In the 1980s, the term picowave was used to describe food irradiation in order to overcome public resistance by playing on the well-known safety of microwave radiation. Find the energy in MeV of a photon having a wavelength of a picometer.

Solution:

1.24 MeV

Exercise:

Problem: Find the mass of ^{239}Pu that has an activity of 1.00 μCi.

Glossary

gray (Gy)
: the SI unit for radiation dose which is defined to be $1\text{ Gy} = 1\text{ J/kg} = 100\text{ rad}$

linear hypothesis
: assumption that risk is directly proportional to risk from high doses

rad
: the ionizing energy deposited per kilogram of tissue

sievert
: the SI equivalent of the rem

relative biological effectiveness (RBE)
: a number that expresses the relative amount of damage that a fixed amount of ionizing radiation of a given type can inflict on biological tissues

quality factor
: same as relative biological effectiveness

roentgen equivalent man (rem)
: a dose unit more closely related to effects in biological tissue

low dose
: a dose less than 100 mSv (10 rem)

moderate dose
: a dose from 0.1 Sv to 1 Sv (10 to 100 rem)

high dose

a dose greater than 1 Sv (100 rem)

hormesis

a term used to describe generally favorable biological responses to low exposures of toxins or radiation

shielding

a technique to limit radiation exposure

Therapeutic Uses of Ionizing Radiation

- Explain the concept of radiotherapy and list typical doses for cancer therapy.

Therapeutic applications of ionizing radiation, called radiation therapy or **radiotherapy**, have existed since the discovery of x-rays and nuclear radioactivity. Today, radiotherapy is used almost exclusively for cancer therapy, where it saves thousands of lives and improves the quality of life and longevity of many it cannot save. Radiotherapy may be used alone or in combination with surgery and chemotherapy (drug treatment) depending on the type of cancer and the response of the patient. A careful examination of all available data has established that radiotherapy's beneficial effects far outweigh its long-term risks.

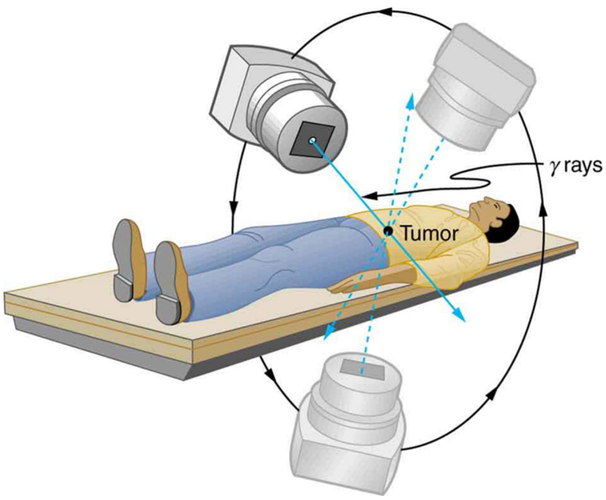
Medical Application

The earliest uses of ionizing radiation on humans were mostly harmful, with many at the level of snake oil as seen in [link]. Radium-doped cosmetics that glowed in the dark were used around the time of World War I. As recently as the 1950s, radon mine tours were promoted as healthful and rejuvenating—those who toured were exposed but gained no benefits. Radium salts were sold as health elixirs for many years. The gruesome death of a wealthy industrialist, who became psychologically addicted to the brew, alerted the unsuspecting to the dangers of radium salt elixirs. Most abuses finally ended after the legislation in the 1950s.

The properties of radiation were once touted for far more than its modern use in cancer therapy. Until 1932, radium was advertised for a variety of uses, often with tragic results. (credit: Struthious Bandersnatch.)

Radiotherapy is effective against cancer because cancer cells reproduce rapidly and, consequently, are more sensitive to radiation. The central problem in radiotherapy is to make the dose for cancer cells as high as possible while limiting the dose for normal cells. The ratio of abnormal cells killed to normal cells killed is called the **therapeutic ratio**, and all radiotherapy techniques are designed to enhance this ratio. Radiation can be concentrated in cancerous tissue by a number of techniques. One of the most prevalent techniques for well-defined tumors is a geometric technique

shown in [link]. A narrow beam of radiation is passed through the patient from a variety of directions with a common crossing point in the tumor. This concentrates the dose in the tumor while spreading it out over a large volume of normal tissue. The external radiation can be x-rays, ^{60}Co γ rays, or ionizing-particle beams produced by accelerators. Accelerator-produced beams of neutrons, π-mesons, and heavy ions such as nitrogen nuclei have been employed, and these can be quite effective. These particles have larger QFs or RBEs and sometimes can be better localized, producing a greater therapeutic ratio. But accelerator radiotherapy is much more expensive and less frequently employed than other forms.



The ^{60}Co source of γ-radiation is rotated around the patient so that the common crossing point is in the tumor, concentrating the dose there. This geometric technique works for well-defined tumors.

Another form of radiotherapy uses chemically inert radioactive implants. One use is for prostate cancer. Radioactive seeds (about 40 to 100 and the size of a grain of rice) are placed in the prostate region. The isotopes used

are usually ^{135}I (6-month half life) or ^{103}Pd (3-month half life). Alpha emitters have the dual advantages of a large QF and a small range for better localization.

Radiopharmaceuticals are used for cancer therapy when they can be localized well enough to produce a favorable therapeutic ratio. Thyroid cancer is commonly treated utilizing radioactive iodine. Thyroid cells concentrate iodine, and cancerous thyroid cells are more aggressive in doing this. An ingenious use of radiopharmaceuticals in cancer therapy tags antibodies with radioisotopes. Antibodies produced by a patient to combat his cancer are extracted, cultured, loaded with a radioisotope, and then returned to the patient. The antibodies are concentrated almost entirely in the tissue they developed to fight, thus localizing the radiation in abnormal tissue. The therapeutic ratio can be quite high for short-range radiation. There is, however, a significant dose for organs that eliminate radiopharmaceuticals from the body, such as the liver, kidneys, and bladder. As with most radiotherapy, the technique is limited by the tolerable amount of damage to the normal tissue.

[link] lists typical therapeutic doses of radiation used against certain cancers. The doses are large, but not fatal because they are localized and spread out in time. Protocols for treatment vary with the type of cancer and the condition and response of the patient. Three to five 200-rem treatments per week for a period of several weeks is typical. Time between treatments allows the body to repair normal tissue. This effect occurs because damage is concentrated in the abnormal tissue, and the abnormal tissue is more sensitive to radiation. Damage to normal tissue limits the doses. You will note that the greatest doses are given to any tissue that is not rapidly reproducing, such as in the adult brain. Lung cancer, on the other end of the scale, cannot ordinarily be cured with radiation because of the sensitivity of lung tissue and blood to radiation. But radiotherapy for lung cancer does alleviate symptoms and prolong life and is therefore justified in some cases.

Type of Cancer	Typical dose (Sv)
Lung	10–20
Hodgkin’s disease	40–45
Skin	40–50
Ovarian	50–75
Breast	50–80+
Brain	80+
Neck	80+
Bone	80+
Soft tissue	80+
Thyroid	80+

Cancer Radiotherapy

Finally, it is interesting to note that chemotherapy employs drugs that interfere with cell division and is, thus, also effective against cancer. It also has almost the same side effects, such as nausea and hair loss, and risks, such as the inducement of another cancer.

Section Summary

- Radiotherapy is the use of ionizing radiation to treat ailments, now limited to cancer therapy.
- The sensitivity of cancer cells to radiation enhances the ratio of cancer cells killed to normal cells killed, which is called the therapeutic ratio.

- Doses for various organs are limited by the tolerance of normal tissue for radiation. Treatment is localized in one region of the body and spread out in time.

Conceptual Questions

Exercise:

Problem:

Radiotherapy is more likely to be used to treat cancer in elderly patients than in young ones. Explain why. Why is radiotherapy used to treat young people at all?

Problems & Exercises

Exercise:

Problem:

A beam of 168-MeV nitrogen nuclei is used for cancer therapy. If this beam is directed onto a 0.200-kg tumor and gives it a 2.00-Sv dose, how many nitrogen nuclei were stopped? (Use an RBE of 20 for heavy ions.)

Solution:

7.44×10^8

Exercise:

Problem:

(a) If the average molecular mass of compounds in food is 50.0 g, how many molecules are there in 1.00 kg of food? (b) How many ion pairs are created in 1.00 kg of food, if it is exposed to 1000 Sv and it takes 32.0 eV to create an ion pair? (c) Find the ratio of ion pairs to molecules. (d) If these ion pairs recombine into a distribution of 2000 new compounds, how many parts per billion is each?

Exercise:

Problem:

Calculate the dose in Sv to the chest of a patient given an x-ray under the following conditions. The x-ray beam intensity is $1.50\ \text{W/m}^2$, the area of the chest exposed is $0.0750\ \text{m}^2$, 35.0% of the x-rays are absorbed in 20.0 kg of tissue, and the exposure time is 0.250 s.

Solution:

$4.92 \times 10^{-4}\ \text{Sv}$

Exercise:

Problem:

(a) A cancer patient is exposed to γ rays from a 5000-Ci ^{60}Co transillumination unit for 32.0 s. The γ rays are collimated in such a manner that only 1.00% of them strike the patient. Of those, 20.0% are absorbed in a tumor having a mass of 1.50 kg. What is the dose in rem to the tumor, if the average γ energy per decay is 1.25 MeV? None of the β s from the decay reach the patient. (b) Is the dose consistent with stated therapeutic doses?

Exercise:

Problem:

What is the mass of ^{60}Co in a cancer therapy transillumination unit containing 5.00 kCi of ^{60}Co?

Solution:

4.43 g

Exercise:

Problem:

Large amounts of $^{65}\mathrm{Zn}$ are produced in copper exposed to accelerator beams. While machining contaminated copper, a physicist ingests $50.0\ \mu\mathrm{Ci}$ of $^{65}\mathrm{Zn}$. Each $^{65}\mathrm{Zn}$ decay emits an average γ-ray energy of 0.550 MeV, 40.0% of which is absorbed in the scientist's 75.0-kg body. What dose in mSv is caused by this in one day?

Exercise:

Problem:

Naturally occurring $^{40}\mathrm{K}$ is listed as responsible for 16 mrem/y of background radiation. Calculate the mass of $^{40}\mathrm{K}$ that must be inside the 55-kg body of a woman to produce this dose. Each $^{40}\mathrm{K}$ decay emits a 1.32-MeV β, and 50% of the energy is absorbed inside the body.

Solution:

0.010 g

Exercise:

Problem:

(a) Background radiation due to $^{226}\mathrm{Ra}$ averages only 0.01 mSv/y, but it can range upward depending on where a person lives. Find the mass of $^{226}\mathrm{Ra}$ in the 80.0-kg body of a man who receives a dose of 2.50-mSv/y from it, noting that each $^{226}\mathrm{Ra}$ decay emits a 4.80-MeV α particle. You may neglect dose due to daughters and assume a constant amount, evenly distributed due to balanced ingestion and bodily elimination. (b) Is it surprising that such a small mass could cause a measurable radiation dose? Explain.

Exercise:

Problem:

The annual radiation dose from ^{14}C in our bodies is 0.01 mSv/y. Each ^{14}C decay emits a β^- averaging 0.0750 MeV. Taking the fraction of ^{14}C to be 1.3×10^{-12} N of normal ^{12}C, and assuming the body is 13% carbon, estimate the fraction of the decay energy absorbed. (The rest escapes, exposing those close to you.)

Solution:

95%

Exercise:

Problem:

If everyone in Australia received an extra 0.05 mSv per year of radiation, what would be the increase in the number of cancer deaths per year? (Assume that time had elapsed for the effects to become apparent.) Assume that there are 200×10^{-4} deaths per Sv of radiation per year. What percent of the actual number of cancer deaths recorded is this?

Glossary

radiotherapy
: the use of ionizing radiation to treat ailments

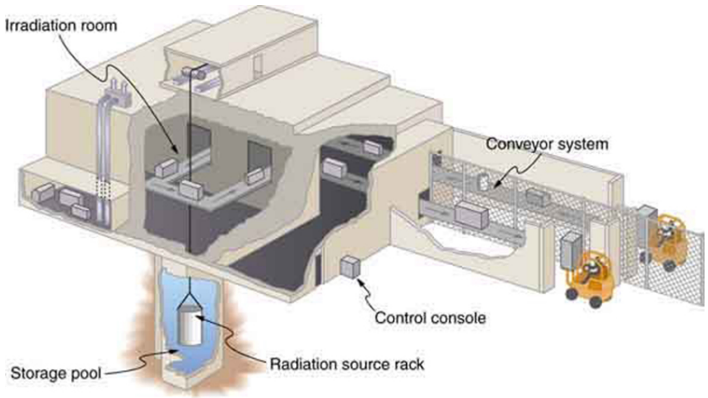
therapeutic ratio
: the ratio of abnormal cells killed to normal cells killed

Food Irradiation

- Define food irradiation low dose, and free radicals.

Ionizing radiation is widely used to sterilize medical supplies, such as bandages, and consumer products, such as tampons. Worldwide, it is also used to irradiate food, an application that promises to grow in the future. **Food irradiation** is the treatment of food with ionizing radiation. It is used to reduce pest infestation and to delay spoilage and prevent illness caused by microorganisms. Food irradiation is controversial. Proponents see it as superior to pasteurization, preservatives, and insecticides, supplanting dangerous chemicals with a more effective process. Opponents see its safety as unproven, perhaps leaving worse toxic residues as well as presenting an environmental hazard at treatment sites. In developing countries, food irradiation might increase crop production by 25.0% or more, and reduce food spoilage by a similar amount. It is used chiefly to treat spices and some fruits, and in some countries, red meat, poultry, and vegetables. Over 40 countries have approved food irradiation at some level.

Food irradiation exposes food to large doses of γ rays, x-rays, or electrons. These photons and electrons induce no nuclear reactions and thus create *no residual radioactivity*. (Some forms of ionizing radiation, such as neutron irradiation, cause residual radioactivity. These are not used for food irradiation.) The γ source is usually ^{60}Co or ^{137}Cs, the latter isotope being a major by-product of nuclear power. Cobalt-60 γ rays average 1.25 MeV, while those of ^{137}Cs are 0.67 MeV and are less penetrating. X-rays used for food irradiation are created with voltages of up to 5 million volts and, thus, have photon energies up to 5 MeV. Electrons used for food irradiation are accelerated to energies up to 10 MeV. The higher the energy per particle, the more penetrating the radiation is and the more ionization it can create. [link] shows a typical γ-irradiation plant.



A food irradiation plant has a conveyor system to pass items through an intense radiation field behind thick shielding walls. The γ source is lowered into a deep pool of water for safe storage when not in use. Exposure times of up to an hour expose food to doses up to 10^4 Gy.

Owing to the fact that food irradiation seeks to destroy organisms such as insects and bacteria, much larger doses than those fatal to humans must be applied. Generally, the simpler the organism, the more radiation it can tolerate. (Cancer cells are a partial exception, because they are rapidly reproducing and, thus, more sensitive.) Current licensing allows up to 1000 Gy to be applied to fresh fruits and vegetables, called a *low dose* in food irradiation. Such a dose is enough to prevent or reduce the growth of many microorganisms, but about 10,000 Gy is needed to kill salmonella, and even more is needed to kill fungi. Doses greater than 10,000 Gy are considered to be high doses in food irradiation and product sterilization.

The effectiveness of food irradiation varies with the type of food. Spices and many fruits and vegetables have dramatically longer shelf lives. These also show no degradation in taste and no loss of food value or vitamins. If not for the mandatory labeling, such foods subjected to low-level irradiation (up to 1000 Gy) could not be distinguished from untreated foods in quality.

However, some foods actually spoil faster after irradiation, particularly those with high water content like lettuce and peaches. Others, such as milk, are given a noticeably unpleasant taste. High-level irradiation produces significant and chemically measurable changes in foods. It produces about a 15% loss of nutrients and a 25% loss of vitamins, as well as some change in taste. Such losses are similar to those that occur in ordinary freezing and cooking.

How does food irradiation work? Ionization produces a random assortment of broken molecules and ions, some with unstable oxygen- or hydrogen-containing molecules known as **free radicals**. These undergo rapid chemical reactions, producing perhaps four or five thousand different compounds called **radiolytic products**, some of which make cell function impossible by breaking cell membranes, fracturing DNA, and so on. How safe is the food afterward? Critics argue that the radiolytic products present a lasting hazard, perhaps being carcinogenic. However, the safety of irradiated food is not known precisely. We do know that low-level food irradiation produces no compounds in amounts that can be measured chemically. This is not surprising, since trace amounts of several thousand compounds may be created. We also know that there have been no observable negative short-term effects on consumers. Long-term effects may show up if large number of people consume large quantities of irradiated food, but no effects have appeared due to the small amounts of irradiated food that are consumed regularly. The case for safety is supported by testing of animal diets that were irradiated; no transmitted genetic effects have been observed. Food irradiation (at least up to a million rad) has been endorsed by the World Health Organization and the UN Food and Agricultural Organization. Finally, the hazard to consumers, if it exists, must be weighed against the benefits in food production and preservation. It must also be weighed against the very real hazards of existing insecticides and food preservatives.

Section Summary

- Food irradiation is the treatment of food with ionizing radiation.
- Irradiating food can destroy insects and bacteria by creating free radicals and radiolytic products that can break apart cell membranes.

- Food irradiation has produced no observable negative short-term effects for humans, but its long-term effects are unknown.

Conceptual Questions

Exercise:

Problem:

Does food irradiation leave the food radioactive? To what extent is the food altered chemically for low and high doses in food irradiation?

Exercise:

Problem:

Compare a low dose of radiation to a human with a low dose of radiation used in food treatment.

Exercise:

Problem:

Suppose one food irradiation plant uses a ^{137}Cs source while another uses an equal activity of ^{60}Co. Assuming equal fractions of the γ rays from the sources are absorbed, why is more time needed to get the same dose using the ^{137}Cs source?

Glossary

food irradiation
: treatment of food with ionizing radiation

free radicals
: ions with unstable oxygen- or hydrogen-containing molecules

radiolytic products
: compounds produced due to chemical reactions of free radicals

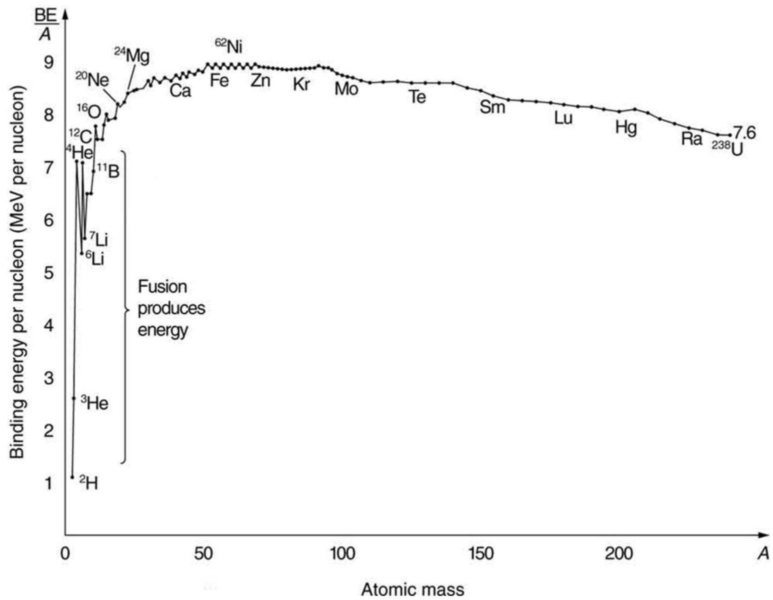
Fusion

- Define nuclear fusion.
- Discuss processes to achieve practical fusion energy generation.

While basking in the warmth of the summer sun, a student reads of the latest breakthrough in achieving sustained thermonuclear power and vaguely recalls hearing about the cold fusion controversy. The three are connected. The Sun’s energy is produced by nuclear fusion (see [link]). Thermonuclear power is the name given to the use of controlled nuclear fusion as an energy source. While research in the area of thermonuclear power is progressing, high temperatures and containment difficulties remain. The cold fusion controversy centered around unsubstantiated claims of practical fusion power at room temperatures.

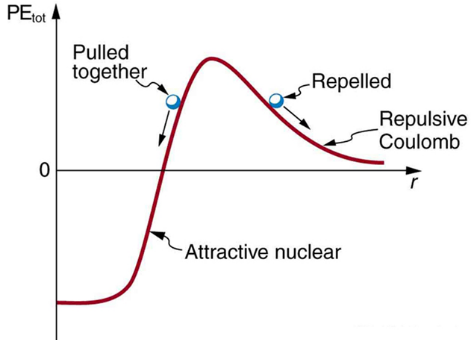
The Sun’s energy is produced by nuclear fusion. (credit: Spiralz)

Nuclear fusion is a reaction in which two nuclei are combined, or *fused*, to form a larger nucleus. We know that all nuclei have less mass than the sum of the masses of the protons and neutrons that form them. The missing mass times c^2 equals the binding energy of the nucleus—the greater the binding energy, the greater the missing mass. We also know that BE/A, the binding energy per nucleon, is greater for medium-mass nuclei and has a maximum at Fe (iron). This means that if two low-mass nuclei can be fused together to form a larger nucleus, energy can be released. The larger nucleus has a greater binding energy and less mass per nucleon than the two that combined. Thus mass is destroyed in the fusion reaction, and energy is released (see [link]). On average, fusion of low-mass nuclei releases energy, but the details depend on the actual nuclides involved.



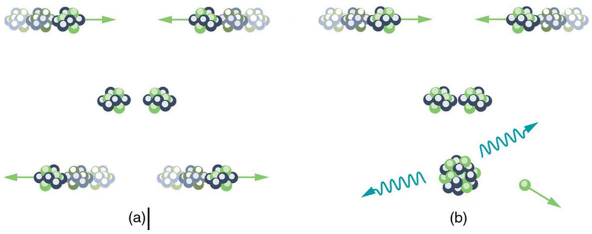
Fusion of light nuclei to form medium-mass nuclei destroys mass, because BE/A is greater for the product nuclei. The larger BE/A is, the less mass per nucleon, and so mass is converted to energy and released in these fusion reactions.

The major obstruction to fusion is the Coulomb repulsion between nuclei. Since the attractive nuclear force that can fuse nuclei together is short ranged, the repulsion of like positive charges must be overcome to get nuclei close enough to induce fusion. [link] shows an approximate graph of the potential energy between two nuclei as a function of the distance between their centers. The graph is analogous to a hill with a well in its center. A ball rolled from the right must have enough kinetic energy to get over the hump before it falls into the deeper well with a net gain in energy. So it is with fusion. If the nuclei are given enough kinetic energy to overcome the electric potential energy due to repulsion, then they can combine, release energy, and fall into a deep well. One way to accomplish this is to heat fusion fuel to high temperatures so that the kinetic energy of thermal motion is sufficient to get the nuclei together.



Potential energy between two light nuclei graphed as a function of distance between them. If the nuclei have enough kinetic energy to get over the Coulomb repulsion hump, they combine, release energy, and drop into a deep attractive well. Tunneling through the barrier is important in practice. The greater the kinetic energy and the higher the particles get up the barrier (or the lower the barrier), the more likely the tunneling.

You might think that, in the core of our Sun, nuclei are coming into contact and fusing. However, in fact, temperatures on the order of $10^8 K$ are needed to actually get the nuclei in contact, exceeding the core temperature of the Sun. Quantum mechanical tunneling is what makes fusion in the Sun possible, and tunneling is an important process in most other practical applications of fusion, too. Since the probability of tunneling is extremely sensitive to barrier height and width, increasing the temperature greatly increases the rate of fusion. The closer reactants get to one another, the more likely they are to fuse (see [link]). Thus most fusion in the Sun and other stars takes place at their centers, where temperatures are highest. Moreover, high temperature is needed for thermonuclear power to be a practical source of energy.



(a) Two nuclei heading toward each other slow down, then stop, and then fly away without touching or fusing. (b) At higher energies, the two nuclei approach close enough for fusion via tunneling. The probability of tunneling increases as they approach,

but they do not have to touch for the reaction to occur.

The Sun produces energy by fusing protons or hydrogen nuclei ^1H (by far the Sun's most abundant nuclide) into helium nuclei ^4He. The principal sequence of fusion reactions forms what is called the **proton-proton cycle**:

Equation:

$$^1\text{H} + {}^1\text{H} \rightarrow {}^2\text{H} + e^+ + v_\text{e} \qquad (0.42\text{ MeV})$$

Equation:

$$^1\text{H} + {}^2\text{H} \rightarrow {}^3\text{He} + \gamma \qquad (5.49\text{ MeV})$$

Equation:

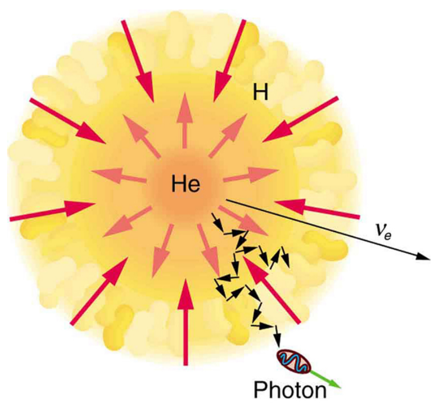
$$^3\text{He} + {}^3\text{He} \rightarrow {}^4\text{He} + {}^1\text{H} + {}^1\text{H} \qquad (12.86\text{ MeV})$$

where e^+ stands for a positron and v_e is an electron neutrino. (The energy in parentheses is *released* by the reaction.) Note that the first two reactions must occur twice for the third to be possible, so that the cycle consumes six protons (^1H) but gives back two. Furthermore, the two positrons produced will find two electrons and annihilate to form four more γ rays, for a total of six. The overall effect of the cycle is thus

Equation:

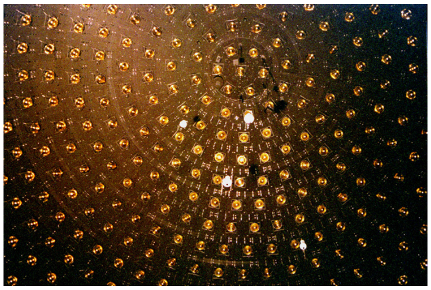
$$2e^- + 4{}^1\text{H} \rightarrow {}^4\text{He} + 2\text{v}_\text{e} + 6\gamma \qquad (26.7\text{ MeV})$$

where the 26.7 MeV includes the annihilation energy of the positrons and electrons and is distributed among all the reaction products. The solar interior is dense, and the reactions occur deep in the Sun where temperatures are highest. It takes about 32,000 years for the energy to diffuse to the surface and radiate away. However, the neutrinos escape the Sun in less than two seconds, carrying their energy with them, because they interact so weakly that the Sun is transparent to them. Negative feedback in the Sun acts as a thermostat to regulate the overall energy output. For instance, if the interior of the Sun becomes hotter than normal, the reaction rate increases, producing energy that expands the interior. This cools it and lowers the reaction rate. Conversely, if the interior becomes too cool, it contracts, increasing the temperature and reaction rate (see [link]). Stars like the Sun are stable for billions of years, until a significant fraction of their hydrogen has been depleted. What happens then is discussed in Introduction to Frontiers of Physics .

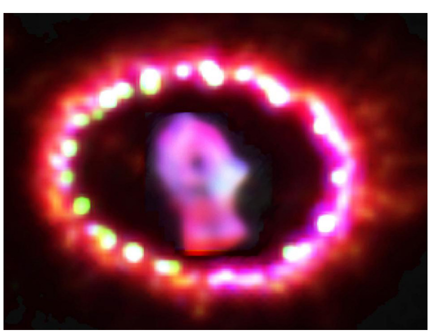


Nuclear fusion in the Sun converts hydrogen nuclei into helium; fusion occurs primarily at the boundary of the helium core, where temperature is highest and sufficient hydrogen remains. Energy released diffuses slowly to the surface, with the exception of neutrinos, which escape immediately. Energy production remains stable because of negative feedback effects.

Theories of the proton-proton cycle (and other energy-producing cycles in stars) were pioneered by the German-born, American physicist Hans Bethe (1906–2005), starting in 1938. He was awarded the 1967 Nobel Prize in physics for this work, and he has made many other contributions to physics and society. Neutrinos produced in these cycles escape so readily that they provide us an excellent means to test these theories and study stellar interiors. Detectors have been constructed and operated for more than four decades now to measure solar neutrinos (see [link]). Although solar neutrinos are detected and neutrinos were observed from Supernova 1987A ([link]), too few solar neutrinos were observed to be consistent with predictions of solar energy production. After many years, this solar neutrino problem was resolved with a blend of theory and experiment that showed that the neutrino does indeed have mass. It was also found that there are three types of neutrinos, each associated with a different type of nuclear decay.



This array of photomultiplier tubes is part of the large solar neutrino detector at the Fermi National Accelerator Laboratory in Illinois. In these experiments, the neutrinos interact with heavy water and produce flashes of light, which are detected by the photomultiplier tubes. In spite of its size and the huge flux of neutrinos that strike it, very few are detected each day since they interact so weakly. This, of course, is the same reason they escape the Sun so readily. (credit: Fred Ullrich)



Supernovas are the source of elements heavier than

iron. Energy released powers nucleosynthesis. Spectroscopic analysis of the ring of material ejected by Supernova 1987A observable in the southern hemisphere, shows evidence of heavy elements. The study of this supernova also provided indications that neutrinos might have mass. (credit: NASA, ESA, and P. Challis)

The proton-proton cycle is not a practical source of energy on Earth, in spite of the great abundance of hydrogen (^1H). The reaction $^1H + {}^1H \rightarrow {}^2H + e^+ + v_e$ has a very low probability of occurring. (This is why our Sun will last for about ten billion years.) However, a number of other fusion reactions are easier to induce. Among them are:
Equation:

$$^2\mathrm{H} + {}^2\mathrm{H} \rightarrow {}^3\mathrm{H} + {}^1\mathrm{H} \qquad (4.03\ \mathrm{MeV})$$

Equation:

$$^2\mathrm{H} + {}^2\mathrm{H} \rightarrow {}^3\mathrm{He} + n \qquad (3.27\ \mathrm{MeV})$$

Equation:

$$^2\mathrm{H} + {}^3\mathrm{H} \rightarrow {}^4\mathrm{He} + n \qquad (17.59\ \mathrm{MeV})$$

Equation:

$$^2\mathrm{H} + {}^2\mathrm{H} \rightarrow {}^4\mathrm{He} + \gamma \qquad (23.85\ \mathrm{MeV}).$$

Deuterium (^2H) is about 0.015% of natural hydrogen, so there is an immense amount of it in sea water alone. In addition to an abundance of deuterium fuel, these fusion reactions produce large energies per reaction (in parentheses), but they do not produce much radioactive waste. Tritium (^3H) is radioactive, but it is consumed as a fuel (the reaction $^2H + {}^3H \rightarrow {}^4He + n$), and the neutrons and γs can be shielded. The neutrons produced can also be used to create more energy and fuel in reactions like
Equation:

$$n + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + \gamma \qquad (20.68\ \text{MeV})$$

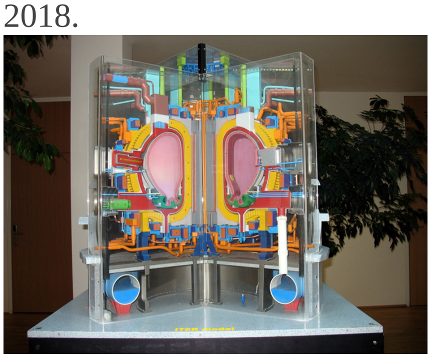
and
Equation:

$$n + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + \gamma \qquad (2.22\ \text{MeV}).$$

Note that these last two reactions, and ${}^{2}\text{H} + {}^{2}\text{H} \rightarrow {}^{4}\text{He} + \gamma$, put most of their energy output into the γ ray, and such energy is difficult to utilize.

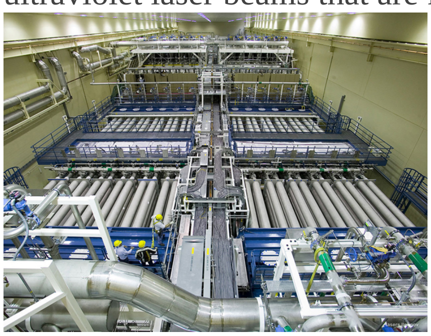
The three keys to practical fusion energy generation are to achieve the temperatures necessary to make the reactions likely, to raise the density of the fuel, and to confine it long enough to produce large amounts of energy. These three factors—temperature, density, and time—complement one another, and so a deficiency in one can be compensated for by the others. **Ignition** is defined to occur when the reactions produce enough energy to be self-sustaining after external energy input is cut off. This goal, which must be reached before commercial plants can be a reality, has not been achieved. Another milestone, called **break-even**, occurs when the fusion power produced equals the heating power input. Break-even has nearly been reached and gives hope that ignition and commercial plants may become a reality in a few decades.

Two techniques have shown considerable promise. The first of these is called **magnetic confinement** and uses the property that charged particles have difficulty crossing magnetic field lines. The tokamak, shown in [link], has shown particular promise. The tokamak's toroidal coil confines charged particles into a circular path with a helical twist due to the circulating ions themselves. In 1995, the Tokamak Fusion Test Reactor at Princeton in the US achieved world-record plasma temperatures as high as 500 million degrees Celsius. This facility operated between 1982 and 1997. A joint international effort is underway in France to build a tokamak-type reactor that will be the stepping stone to commercial power. ITER, as it is called, will be a full-scale device that aims to demonstrate the feasibility of fusion energy. It will generate 500 MW of power for extended periods of time and will achieve break-even conditions. It will study plasmas in conditions similar to those expected in a fusion power plant. Completion is scheduled for 2018.



(a) Artist's rendition of ITER, a tokamak-type fusion reactor being built in southern France. It is hoped that this gigantic machine will reach the break-even point. Completion is scheduled for 2018. (credit: Stephan Mosel, Flickr)

The second promising technique aims multiple lasers at tiny fuel pellets filled with a mixture of deuterium and tritium. Huge power input heats the fuel, evaporating the confining pellet and crushing the fuel to high density with the expanding hot plasma produced. This technique is called **inertial confinement**, because the fuel's inertia prevents it from escaping before significant fusion can take place. Higher densities have been reached than with tokamaks, but with smaller confinement times. In 2009, the Lawrence Livermore Laboratory (CA) completed a laser fusion device with 192 ultraviolet laser beams that are focused upon a D-T pellet (see [link]).



National Ignition Facility (CA). This image shows a laser bay where 192 laser beams will focus onto a small D-T target, producing fusion. (credit: Lawrence Livermore National Laboratory, Lawrence Livermore National Security, LLC, and the Department of Energy)

Example:
Calculating Energy and Power from Fusion

(a) Calculate the energy released by the fusion of a 1.00-kg mixture of deuterium and tritium, which produces helium. There are equal numbers of deuterium and tritium nuclei in the mixture.

(b) If this takes place continuously over a period of a year, what is the average power output?

Strategy

According to ${}^{2}\text{H} + {}^{3}\text{H} \rightarrow {}^{4}\text{He} + n$, the energy per reaction is 17.59 MeV. To find the total energy released, we must find the number of deuterium and tritium atoms in a kilogram. Deuterium has an atomic mass of about 2 and tritium has an atomic mass of about 3, for a total of about 5 g per mole of reactants or about 200 mol in 1.00 kg. To get a more precise figure, we will use the atomic masses from Appendix A. The power output is best expressed in watts, and so the energy output needs to be calculated in joules and then divided by the number of seconds in a year.

Solution for (a)

The atomic mass of deuterium (${}^{2}\text{H}$) is 2.014102 u, while that of tritium (${}^{3}\text{H}$) is 3.016049 u, for a total of 5.032151 u per reaction. So a mole of reactants has a mass of 5.03 g, and in 1.00 kg there are $(1000\text{ g})/(5.03\text{ g/mol}){=}198.8\text{ mol of reactants}$. The number of reactions that take place is therefore

Equation:

$$(198.8\text{ mol})\left(6.02 \times 10^{23}\text{ mol}^{-1}\right) = 1.20 \times 10^{26}\text{ reactions.}$$

The total energy output is the number of reactions times the energy per reaction:

Equation:

$$\begin{aligned} E = \left(1.20 \times 10^{26}\text{ reactions}\right)(17.59\text{ MeV/reaction})\left(1.602 \times 10^{-13}\text{ J/MeV}\right) \\ = 3.37 \times 10^{14}\text{ J.} \end{aligned}$$

Solution for (b)

Power is energy per unit time. One year has 3.16×10^{7} s, so

Equation:

$$\begin{aligned} P &= \frac{E}{t} = \frac{3.37 \times 10^{14}\text{ J}}{3.16 \times 10^{7}\text{ s}} \\ &= 1.07 \times 10^{7}\text{ W} = 10.7\text{ MW.} \end{aligned}$$

Discussion

By now we expect nuclear processes to yield large amounts of energy, and we are not disappointed here. The energy output of 3.37×10^{14} J from fusing 1.00 kg of deuterium

and tritium is equivalent to 2.6 million gallons of gasoline and about eight times the energy output of the bomb that destroyed Hiroshima. Yet the average backyard swimming pool has about 6 kg of deuterium in it, so that fuel is plentiful if it can be utilized in a controlled manner. The average power output over a year is more than 10 MW, impressive but a bit small for a commercial power plant. About 32 times this power output would allow generation of 100 MW of electricity, assuming an efficiency of one-third in converting the fusion energy to electrical energy.

Section Summary

- Nuclear fusion is a reaction in which two nuclei are combined to form a larger nucleus. It releases energy when light nuclei are fused to form medium-mass nuclei.
- Fusion is the source of energy in stars, with the proton-proton cycle,
  Equation:
  $$^{1}\text{H} + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + e^{+} + v_{\text{e}} \qquad (0.42\text{ MeV})$$
  Equation:
  $$^{1}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + \gamma \qquad (5.49\text{ MeV})$$
  Equation:
  $$^{3}\text{He} + {}^{3}\text{He} \rightarrow {}^{4}\text{He} + {}^{1}\text{H} + {}^{1}\text{H} \qquad (12.86\text{ MeV})$$
  being the principal sequence of energy-producing reactions in our Sun.
- The overall effect of the proton-proton cycle is
  Equation:
  $$2e^{-} + 4{}^{1}\text{H} \rightarrow {}^{4}\text{He} + 2v_{\text{e}} + 6\gamma \qquad (26.7\text{ MeV}),$$
  where the 26.7 MeV includes the energy of the positrons emitted and annihilated.
- Attempts to utilize controlled fusion as an energy source on Earth are related to deuterium and tritium, and the reactions play important roles.
- Ignition is the condition under which controlled fusion is self-sustaining; it has not yet been achieved. Break-even, in which the fusion energy output is as great as the external energy input, has nearly been achieved.
- Magnetic confinement and inertial confinement are the two methods being developed for heating fuel to sufficiently high temperatures, at sufficient density, and for sufficiently long times to achieve ignition. The first method uses magnetic fields

and the second method uses the momentum of impinging laser beams for confinement.

Conceptual Questions

Exercise:

Problem: Why does the fusion of light nuclei into heavier nuclei release energy?

Exercise:

Problem:

Energy input is required to fuse medium-mass nuclei, such as iron or cobalt, into more massive nuclei. Explain why.

Exercise:

Problem:

In considering potential fusion reactions, what is the advantage of the reaction $^{2}\text{H} + {}^{3}\text{H} \rightarrow {}^{4}\text{He} + n$ over the reaction $^{2}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + n$?

Exercise:

Problem:

Give reasons justifying the contention made in the text that energy from the fusion reaction $^{2}\text{H} + {}^{2}\text{H} \rightarrow {}^{4}\text{He} + \gamma$ is relatively difficult to capture and utilize.

Problems & Exercises

Exercise:

Problem:

Verify that the total number of nucleons, total charge, and electron family number are conserved for each of the fusion reactions in the proton-proton cycle in

Equation:

$$^{1}\text{H} + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + e^{+} + v_{\text{e}},$$

Equation:

$$^{1}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + \gamma,$$

and

Equation:

$$ {}^{3}\text{He} + {}^{3}\text{He} \rightarrow {}^{4}\text{He} + {}^{1}\text{H} + {}^{1}\text{H}. $$

(List the value of each of the conserved quantities before and after each of the reactions.)

Solution:

(a) A=1+1=2, Z=1+1=1+1, efn $= 0 = -1 + 1$

(b) A=1+2=3, Z=1+1=2, efn=0=0

(c) A=3+3=4+1+1, Z=2+2=2+1+1, efn=0=0

Exercise:

Problem:

Calculate the energy output in each of the fusion reactions in the proton-proton cycle, and verify the values given in the above summary.

Exercise:

Problem:

Show that the total energy released in the proton-proton cycle is 26.7 MeV, considering the overall effect in ${}^{1}\text{H} + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + e^{+} + v_{\text{e}}$, ${}^{1}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + \gamma$, and ${}^{3}\text{He} + {}^{3}\text{He} \rightarrow {}^{4}\text{He} + {}^{1}\text{H} + {}^{1}\text{H}$ and being certain to include the annihilation energy.

Solution:

$$ \begin{aligned} E &= (m_{\text{i}} - m_{\text{f}})c^2 \\ &= \left[4m\left({}^{1}\text{H}\right) - m\left({}^{4}\text{He}\right)\right]c^2 \\ &= [4(1.007825) - 4.002603](931.5\text{ MeV}) \\ &= 26.73\text{ MeV} \end{aligned} $$

Exercise:

Problem:

Verify by listing the number of nucleons, total charge, and electron family number before and after the cycle that these quantities are conserved in the overall proton-proton cycle in $2e^{-} + 4{}^{1}\text{H} \rightarrow {}^{4}\text{He} + 2v_{\text{e}} + 6\gamma$.

Exercise:

Problem:

The energy produced by the fusion of a 1.00-kg mixture of deuterium and tritium was found in Example Calculating Energy and Power from Fusion. Approximately how many kilograms would be required to supply the annual energy use in the United States?

Solution:

3.12×10^5 kg (about 200 tons)

Exercise:

Problem:

Tritium is naturally rare, but can be produced by the reaction $n + {}^2\text{H} \rightarrow {}^3\text{H} + \gamma$. How much energy in MeV is released in this neutron capture?

Exercise:

Problem: Two fusion reactions mentioned in the text are

$n + {}^3\text{He} \rightarrow {}^4\text{He} + \gamma$

and

$n + {}^1\text{H} \rightarrow {}^2\text{H} + \gamma.$

Both reactions release energy, but the second also creates more fuel. Confirm that the energies produced in the reactions are 20.58 and 2.22 MeV, respectively. Comment on which product nuclide is most tightly bound, ${}^4\text{He}$ or ${}^2\text{H}$.

Solution:

$$\begin{aligned} E &= (m_{\text{i}} - m_{\text{f}})c^2 \\ E_1 &= (1.008665 + 3.016030 - 4.002603)(931.5\text{ MeV}) \\ &= 20.58\text{ MeV} \\ E_2 &= (1.008665 + 1.007825 - 2.014102)(931.5\text{ MeV}) \\ &= 2.224\text{ MeV} \end{aligned}$$

${}^4\text{He}$ is more tightly bound, since this reaction gives off more energy per nucleon.

Exercise:

Problem:

(a) Calculate the number of grams of deuterium in an 80,000-L swimming pool, given deuterium is 0.0150% of natural hydrogen.

(b) Find the energy released in joules if this deuterium is fused via the reaction $^{2}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + n$.

(c) Could the neutrons be used to create more energy?

(d) Discuss the amount of this type of energy in a swimming pool as compared to that in, say, a gallon of gasoline, also taking into consideration that water is far more abundant.

Exercise:

Problem:

How many kilograms of water are needed to obtain the 198.8 mol of deuterium, assuming that deuterium is 0.01500% (by number) of natural hydrogen?

Solution:

1.19×10^{4} kg

Exercise:

Problem: The power output of the Sun is 4×10^{26} W.

(a) If 90% of this is supplied by the proton-proton cycle, how many protons are consumed per second?

(b) How many neutrinos per second should there be per square meter at the Earth from this process? This huge number is indicative of how rarely a neutrino interacts, since large detectors observe very few per day.

Exercise:

Problem:

Another set of reactions that result in the fusing of hydrogen into helium in the Sun and especially in hotter stars is called the carbon cycle. It is

Equation:

$$\begin{array}{lcl} {}^{12}\text{C} + {}^{1}\text{H} & \rightarrow & {}^{13}\text{N} + \gamma, \\ {}^{13}\text{N} & \rightarrow & {}^{13}\text{C} + e^{+} + v_e, \\ {}^{13}\text{C} + {}^{1}\text{H} & \rightarrow & {}^{14}\text{N} + \gamma, \\ {}^{14}\text{N} + {}^{1}\text{H} & \rightarrow & {}^{15}\text{O} + \gamma, \\ {}^{15}\text{O} & \rightarrow & {}^{15}\text{N} + e^{+} + v_e, \\ {}^{15}\text{N} + {}^{1}\text{H} & \rightarrow & {}^{12}\text{C} + {}^{4}\text{He}. \end{array}$$

Write down the overall effect of the carbon cycle (as was done for the proton-proton cycle in $2e^{-} + 4^{1}\text{H} \rightarrow {}^{4}\text{He} + 2v_e + 6\gamma$). Note the number of protons (${}^{1}\text{H}$) required and assume that the positrons (e^{+}) annihilate electrons to form more γ rays.

Solution:

$2e^{-} + 4^{1}\text{H} \rightarrow {}^{4}\text{He} + 7\gamma + 2v_e$

Exercise:

Problem:

(a) Find the total energy released in MeV in each carbon cycle (elaborated in the above problem) including the annihilation energy.

(b) How does this compare with the proton-proton cycle output?

Exercise:

Problem:

Verify that the total number of nucleons, total charge, and electron family number are conserved for each of the fusion reactions in the carbon cycle given in the above problem. (List the value of each of the conserved quantities before and after each of the reactions.)

Solution:

(a) A=12+1=13, Z=6+1=7, efn $= 0 = 0$

(b) A=13=13, Z=7=6+1, efn $= 0 = -1 + 1$

(c) A=13 + 1=14, Z=6+1=7, efn $= 0 = 0$

(d) A=14 + 1=15, Z=7+1=8, efn $= 0 = 0$

(e) A=15=15, Z=8=7+1, efn $= 0 = -1 + 1$

(f) $A{=}15+1{=}12+4$, $Z{=}7{+}1{=}6+2$, $\text{efn}=0=0$

Exercise:

Problem: Integrated Concepts

The laser system tested for inertial confinement can produce a 100-kJ pulse only 1.00 ns in duration. (a) What is the power output of the laser system during the brief pulse?

(b) How many photons are in the pulse, given their wavelength is 1.06 µm?

(c) What is the total momentum of all these photons?

(d) How does the total photon momentum compare with that of a single 1.00 MeV deuterium nucleus?

Exercise:

Problem: Integrated Concepts

Find the amount of energy given to the ^{4}He nucleus and to the γ ray in the reaction $n+^3\text{He}\rightarrow^4\text{He}+\gamma$, using the conservation of momentum principle and taking the reactants to be initially at rest. This should confirm the contention that most of the energy goes to the γ ray.

Solution:

$E_\gamma = 20.6\ \text{MeV}$

$E_{^4\text{He}} = 5.68\times10^{-2}\ \text{MeV}$

Exercise:

Problem: Integrated Concepts

(a) What temperature gas would have atoms moving fast enough to bring two ^{3}He nuclei into contact? Note that, because both are moving, the average kinetic energy only needs to be half the electric potential energy of these doubly charged nuclei when just in contact with one another.

(b) Does this high temperature imply practical difficulties for doing this in controlled fusion?

Exercise:

Problem: **Integrated Concepts**

(a) Estimate the years that the deuterium fuel in the oceans could supply the energy needs of the world. Assume world energy consumption to be ten times that of the United States which is 8×10^{19} J/y and that the deuterium in the oceans could be converted to energy with an efficiency of 32%. You must estimate or look up the amount of water in the oceans and take the deuterium content to be 0.015% of natural hydrogen to find the mass of deuterium available. Note that approximate energy yield of deuterium is 3.37×10^{14} J/kg.

(b) Comment on how much time this is by any human measure. (It is not an unreasonable result, only an impressive one.)

Solution:

(a) 3×10^{9} y

(b) This is approximately half the lifetime of the Earth.

Glossary

break-even
: when fusion power produced equals the heating power input

ignition
: when a fusion reaction produces enough energy to be self-sustaining after external energy input is cut off

inertial confinement
: a technique that aims multiple lasers at tiny fuel pellets evaporating and crushing them to high density

magnetic confinement
: a technique in which charged particles are trapped in a small region because of difficulty in crossing magnetic field lines

nuclear fusion
: a reaction in which two nuclei are combined, or fused, to form a larger nucleus

proton-proton cycle
: the combined reactions $^{1}H + ^{1}H \rightarrow ^{2}H + e^{+} + v_{e}$, $^{1}H + ^{2}H \rightarrow ^{3}He + \gamma$, and $^{3}He + ^{3}He \rightarrow ^{4}He + ^{1}H + ^{1}H$

Fission

- Define nuclear fission.
- Discuss how fission fuel reacts and describe what it produces.
- Describe controlled and uncontrolled chain reactions.

Nuclear fission is a reaction in which a nucleus is split (or *fissured*). Controlled fission is a reality, whereas controlled fusion is a hope for the future. Hundreds of nuclear fission power plants around the world attest to the fact that controlled fission is practical and, at least in the short term, economical, as seen in [link]. Whereas nuclear power was of little interest for decades following TMI and Chernobyl (and now Fukushima Daiichi), growing concerns over global warming has brought nuclear power back on the table as a viable energy alternative. By the end of 2009, there were 442 reactors operating in 30 countries, providing 15% of the world's electricity. France provides over 75% of its electricity with nuclear power, while the US has 104 operating reactors providing 20% of its electricity. Australia and New Zealand have none. China is building nuclear power plants at the rate of one start every month.

The people living near this nuclear power plant have no measurable exposure to radiation that is traceable to the plant. About 16% of the world's electrical power is generated by controlled nuclear fission in such plants. The cooling towers are the most prominent features but are not unique to nuclear

power. The reactor is in the small domed building to the left of the towers. (credit: Kalmthouts)

Fission is the opposite of fusion and releases energy only when heavy nuclei are split. As noted in Fusion, energy is released if the products of a nuclear reaction have a greater binding energy per nucleon (BE/A) than the parent nuclei. [link] shows that BE/A is greater for medium-mass nuclei than heavy nuclei, implying that when a heavy nucleus is split, the products have less mass per nucleon, so that mass is destroyed and energy is released in the reaction. The amount of energy per fission reaction can be large, even by nuclear standards. The graph in [link] shows BE/A to be about 7.6 MeV/nucleon for the heaviest nuclei (A about 240), while BE/A is about 8.6 MeV/nucleon for nuclei having A about 120. Thus, if a heavy nucleus splits in half, then about 1 MeV per nucleon, or approximately 240 MeV per fission, is released. This is about 10 times the energy per fusion reaction, and about 100 times the energy of the average α, β, or γ decay.

Example:
Calculating Energy Released by Fission
Calculate the energy released in the following spontaneous fission reaction:
Equation:

$$^{238}\mathrm{U} \rightarrow {}^{95}\mathrm{Sr} + {}^{140}\mathrm{Xe} + 3n$$

given the atomic masses to be $m(^{238}\mathrm{U}) = 238.050784\ \mathrm{u}$, $m(^{95}\mathrm{Sr}) = 94.919388\ \mathrm{u}$, $m(^{140}\mathrm{Xe}) = 139.921610\ \mathrm{u}$, and $m(n) = 1.008665\ \mathrm{u}$.

Strategy
As always, the energy released is equal to the mass destroyed times c^2 , so we must find the difference in mass between the parent $^{238}\mathrm{U}$ and the fission products.

Solution
The products have a total mass of
Equation:

$$\begin{aligned} m_{\text{products}} &= 94.919388 \text{ u} + 139.921610 \text{ u} + 3(1.008665 \text{ u}) \\ &= 237.866993 \text{ u}. \end{aligned}$$

The mass lost is the mass of ${}^{238}\text{U}$ minus m_{products}, or
Equation:

$$\Delta m = 238.050784 \text{ u} - 237.8669933 \text{ u} = 0.183791 \text{ u},$$

so the energy released is
Equation:

$$\begin{aligned} E &= (\Delta m)c^2 \\ &= (0.183791 \text{ u})\frac{931.5 \text{ MeV}/c^2}{\text{u}}c^2 = 171.2 \text{ MeV}. \end{aligned}$$

Discussion
A number of important things arise in this example. The 171-MeV energy released is large, but a little less than the earlier estimated 240 MeV. This is because this fission reaction produces neutrons and does not split the nucleus into two equal parts. Fission of a given nuclide, such as ${}^{238}\text{U}$, does not always produce the same products. Fission is a statistical process in which an entire range of products are produced with various probabilities. Most fission produces neutrons, although the number varies with each fission. This is an extremely important aspect of fission, because *neutrons can induce more fission*, enabling self-sustaining chain reactions.

Spontaneous fission can occur, but this is usually not the most common decay mode for a given nuclide. For example, ${}^{238}\text{U}$ can spontaneously fission, but it decays mostly by α emission. Neutron-induced fission is crucial as seen in [link]. Being chargeless, even low-energy neutrons can strike a nucleus and be absorbed once they feel the attractive nuclear force. Large nuclei are described by a **liquid drop model** with surface tension and oscillation modes, because the large number of nucleons act like atoms in a drop. The neutron is attracted and thus, deposits energy, causing the nucleus to deform as a liquid drop. If stretched enough, the nucleus narrows in the middle. The number of nucleons in contact and the strength of the nuclear force binding the nucleus together are reduced. Coulomb repulsion between the two ends then succeeds in fissioning the nucleus, which pops like a water drop into two large pieces and a few neutrons. **Neutron-induced fission** can be written as

Equation:

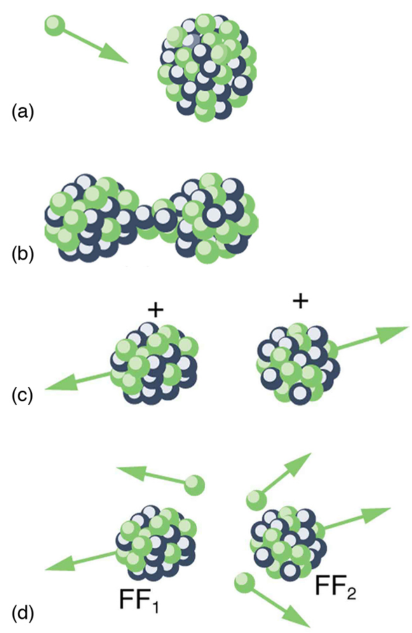
$$n+\ {}^{A}\mathrm{X}\rightarrow\mathrm{FF}_1+\mathrm{FF}_2+\mathrm{xn},$$

where FF_1 and FF_2 are the two daughter nuclei, called **fission fragments**, and x is the number of neutrons produced. Most often, the masses of the fission fragments are not the same. Most of the released energy goes into the kinetic energy of the fission fragments, with the remainder going into the neutrons and excited states of the fragments. Since neutrons can induce fission, a self-sustaining chain reaction is possible, provided more than one neutron is produced on average — that is, if $x>1$ in $n+\ {}^{A}\mathrm{X}\rightarrow\mathrm{FF}_1+\mathrm{FF}_2+\mathrm{xn}$. This can also be seen in [link].

An example of a typical neutron-induced fission reaction is
Equation:

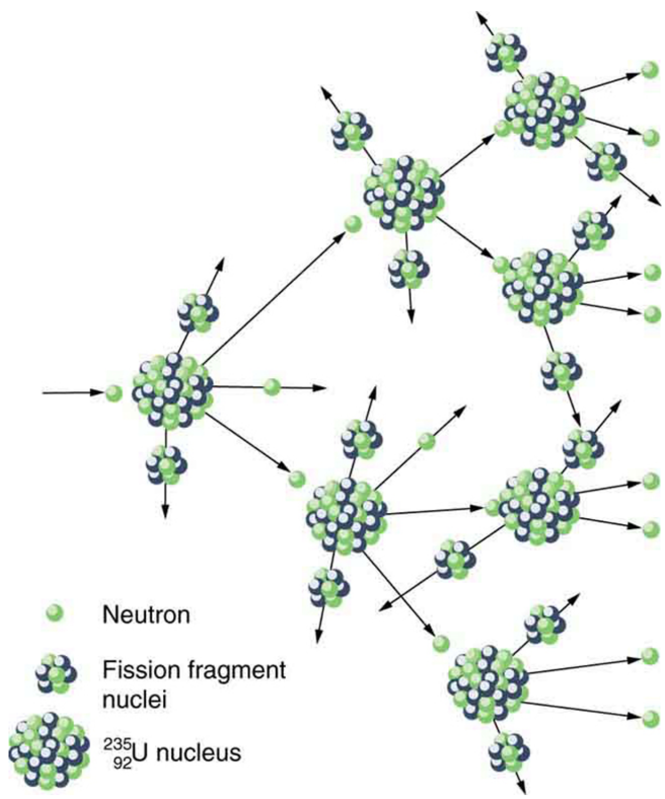
$$n+{}^{235}_{92}\mathrm{U}\rightarrow{}^{142}_{56}\mathrm{Ba}+{}^{91}_{36}\mathrm{Kr}+3n.$$

Note that in this equation, the total charge remains the same (is conserved): $92+0=56+36$. Also, as far as whole numbers are concerned, the mass is constant: $1+235=142+91+3$. This is not true when we consider the masses out to 6 or 7 significant places, as in the previous example.



Neutron-induced

fission is shown. First, energy is put into this large nucleus when it absorbs a neutron. Acting like a struck liquid drop, the nucleus deforms and begins to narrow in the middle. Since fewer nucleons are in contact, the repulsive Coulomb force is able to break the nucleus into two parts with some neutrons also flying away.



A chain reaction can produce self-sustained fission if each fission

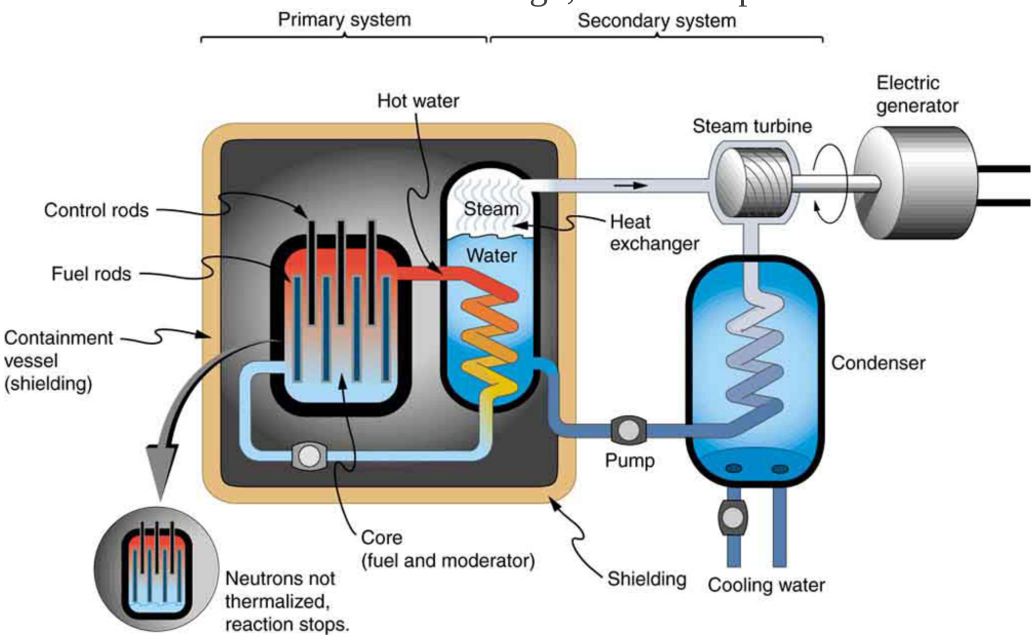
produces enough neutrons to induce at least one more fission. This depends on several factors, including how many neutrons are produced in an average fission and how easy it is to make a particular type of nuclide fission.

Not every neutron produced by fission induces fission. Some neutrons escape the fissionable material, while others interact with a nucleus without making it fission. We can enhance the number of fissions produced by neutrons by having a large amount of fissionable material. The minimum amount necessary for self-sustained fission of a given nuclide is called its **critical mass**. Some nuclides, such as $^{239}\mathrm{Pu}$, produce more neutrons per fission than others, such as $^{235}\mathrm{U}$. Additionally, some nuclides are easier to make fission than others. In particular, $^{235}\mathrm{U}$ and $^{239}\mathrm{Pu}$ are easier to fission than the much more abundant $^{238}\mathrm{U}$. Both factors affect critical mass, which is smallest for $^{239}\mathrm{Pu}$.

The reason $^{235}\mathrm{U}$ and $^{239}\mathrm{Pu}$ are easier to fission than $^{238}\mathrm{U}$ is that the nuclear force is more attractive for an even number of neutrons in a nucleus than for an odd number. Consider that $^{235}_{92}\mathrm{U}_{143}$ has 143 neutrons, and $^{239}_{94}\mathrm{P}_{145}$ has 145 neutrons, whereas $^{238}_{92}\mathrm{U}_{146}$ has 146. When a neutron encounters a nucleus with an odd number of neutrons, the nuclear force is more attractive, because the additional neutron will make the number even. About 2-MeV more energy is deposited in the resulting nucleus than would be the case if the number of neutrons was already even. This extra energy produces greater deformation, making fission more likely. Thus, $^{235}\mathrm{U}$ and $^{239}\mathrm{Pu}$ are superior fission fuels. The isotope $^{235}\mathrm{U}$ is only 0.72 % of natural uranium, while $^{238}\mathrm{U}$ is 99.27%, and $^{239}\mathrm{Pu}$ does not exist in nature. Australia has the largest deposits of uranium in the world, standing at 28% of the total. This is followed by Kazakhstan and Canada. The US has only 3% of global reserves.

Most fission reactors utilize $^{235}\mathrm{U}$, which is separated from $^{238}\mathrm{U}$ at some expense. This is called enrichment. The most common separation method is gaseous diffusion of uranium hexafluoride (UF_6) through membranes. Since $^{235}\mathrm{U}$ has less mass than $^{238}\mathrm{U}$, its UF_6 molecules have higher average velocity at the same temperature and diffuse faster. Another interesting characteristic of $^{235}\mathrm{U}$ is that it preferentially absorbs very slow moving neutrons (with energies a

fraction of an eV), whereas fission reactions produce fast neutrons with energies in the order of an MeV. To make a self-sustained fission reactor with ^{235}U , it is thus necessary to slow down (“thermalize”) the neutrons. Water is very effective, since neutrons collide with protons in water molecules and lose energy. [link] shows a schematic of a reactor design, called the pressurized water reactor.



A pressurized water reactor is cleverly designed to control the fission of large amounts of ^{235}U , while using the heat produced in the fission reaction to create steam for generating electrical energy. Control rods adjust neutron flux so that criticality is obtained, but not exceeded. In case the reactor overheats and boils the water away, the chain reaction terminates, because water is needed to thermalize the neutrons. This inherent safety feature can be overwhelmed in extreme circumstances.

Control rods containing nuclides that very strongly absorb neutrons are used to adjust neutron flux. To produce large power, reactors contain hundreds to thousands of critical masses, and the chain reaction easily becomes self-sustaining, a condition called **criticality**. Neutron flux should be carefully regulated to avoid an exponential increase in fissions, a condition called **supercriticality**. Control rods help prevent overheating, perhaps even a meltdown or explosive disassembly. The water that is used to thermalize

neutrons, necessary to get them to induce fission in ^{235}U , and achieve criticality, provides a negative feedback for temperature increases. In case the reactor overheats and boils the water to steam or is breached, the absence of water kills the chain reaction. Considerable heat, however, can still be generated by the reactor's radioactive fission products. Other safety features, thus, need to be incorporated in the event of a *loss of coolant* accident, including auxiliary cooling water and pumps.

Example:

Calculating Energy from a Kilogram of Fissionable Fuel

Calculate the amount of energy produced by the fission of 1.00 kg of ^{235}U , given the average fission reaction of ^{235}U produces 200 MeV.

Strategy

The total energy produced is the number of ^{235}U atoms times the given energy per ^{235}U fission. We should therefore find the number of ^{235}U atoms in 1.00 kg.

Solution

The number of ^{235}U atoms in 1.00 kg is Avogadro's number times the number of moles. One mole of ^{235}U has a mass of 235.04 g; thus, there are $(1000\text{ g})/(235.04\text{ g/mol}) = 4.25\text{ mol}$. The number of ^{235}U atoms is therefore,

Equation:

$$(4.25\text{ mol})\left(6.02\times10^{23}\ ^{235}\text{U/mol}\right) = 2.56\times10^{24}\ ^{235}\text{U}.$$

So the total energy released is

Equation:

$$\begin{aligned} E &= \left(2.56\times10^{24}\ ^{235}\text{U}\right)\left(\frac{200\text{ MeV}}{^{235}\text{U}}\right)\left(\frac{1.60\times10^{-13}\text{ J}}{\text{MeV}}\right) \\ &= 8.21\times10^{13}\text{ J}. \end{aligned}$$

Discussion

This is another impressively large amount of energy, equivalent to about 14,000 barrels of crude oil or 600,000 gallons of gasoline. But, it is only one-fourth the energy produced by the fusion of a kilogram mixture of deuterium and tritium as seen in [link]. Even though each fission reaction yields about ten times the energy of a fusion reaction, the energy per kilogram of fission fuel is less, because there are far fewer moles per kilogram of the heavy nuclides. Fission

fuel is also much more scarce than fusion fuel, and less than 1% of uranium $(\mathrm{the}\ ^{235}\mathrm{U})$ is readily usable.

One nuclide already mentioned is $^{239}\mathrm{Pu}$, which has a 24,120-y half-life and does not exist in nature. Plutonium-239 is manufactured from $^{238}\mathrm{U}$ in reactors, and it provides an opportunity to utilize the other 99% of natural uranium as an energy source. The following reaction sequence, called **breeding**, produces $^{239}\mathrm{Pu}$. Breeding begins with neutron capture by $^{238}\mathrm{U}$:

Equation:

$$^{238}\mathrm{U}+n\rightarrow\ ^{239}\mathrm{U}+\gamma.$$

Uranium-239 then β^- decays:

Equation:

$$^{239}\mathrm{U}\rightarrow\ ^{239}\mathrm{Np}+\beta^-+v_e(\mathrm{t}_{1/2}=23\ \mathrm{min}).$$

Neptunium-239 also β^- decays:

Equation:

$$^{239}\mathrm{Np}\rightarrow\ ^{239}\mathrm{Pu}+\beta^-+v_e(\mathrm{t}_{1/2}=2.4\ \mathrm{d}).$$

Plutonium-239 builds up in reactor fuel at a rate that depends on the probability of neutron capture by $^{238}\mathrm{U}$ (all reactor fuel contains more $^{238}\mathrm{U}$ than $^{235}\mathrm{U}$). Reactors designed specifically to make plutonium are called **breeder reactors**. They seem to be inherently more hazardous than conventional reactors, but it remains unknown whether their hazards can be made economically acceptable. The four reactors at Chernobyl, including the one that was destroyed, were built to breed plutonium and produce electricity. These reactors had a design that was significantly different from the pressurized water reactor illustrated above.

Plutonium-239 has advantages over $^{235}\mathrm{U}$ as a reactor fuel — it produces more neutrons per fission on average, and it is easier for a thermal neutron to cause it to fission. It is also chemically different from uranium, so it is inherently easier to separate from uranium ore. This means $^{239}\mathrm{Pu}$ has a particularly small critical mass, an advantage for nuclear weapons.

Note:
PhET Explorations: Nuclear Fission
Start a chain reaction, or introduce non-radioactive isotopes to prevent one. Control energy production in a nuclear reactor!

https://archive.cnx.org/specials/01caf0d0-116f-11e6-b891-abfdaa77b03b/nuclear-fission/#sim-one-nucleus

Section Summary

- Nuclear fission is a reaction in which a nucleus is split.
- Fission releases energy when heavy nuclei are split into medium-mass nuclei.
- Self-sustained fission is possible, because neutron-induced fission also produces neutrons that can induce other fissions, $n + {}^{A}X \rightarrow \mathrm{FF}_1 + \mathrm{FF}_2 + \mathrm{xn}$, where FF_1 and FF_2 are the two daughter nuclei, or fission fragments, and x is the number of neutrons produced.
- A minimum mass, called the critical mass, should be present to achieve criticality.
- More than a critical mass can produce supercriticality.
- The production of new or different isotopes (especially ${}^{239}\mathrm{Pu}$) by nuclear transformation is called breeding, and reactors designed for this purpose are called breeder reactors.

Conceptual Questions

Exercise:

Problem:

Explain why the fission of heavy nuclei releases energy. Similarly, why is it that energy input is required to fission light nuclei?

Exercise:

Problem:

Explain, in terms of conservation of momentum and energy, why collisions of neutrons with protons will thermalize neutrons better than collisions with oxygen.

Exercise:

Problem:

The ruins of the Chernobyl reactor are enclosed in a huge concrete structure built around it after the accident. Some rain penetrates the building in winter, and radioactivity from the building increases. What does this imply is happening inside?

Exercise:

Problem:

Since the uranium or plutonium nucleus fissions into several fission fragments whose mass distribution covers a wide range of pieces, would you expect more residual radioactivity from fission than fusion? Explain.

Exercise:

Problem:

The core of a nuclear reactor generates a large amount of thermal energy from the decay of fission products, even when the power-producing fission chain reaction is turned off. Would this residual heat be greatest after the reactor has run for a long time or short time? What if the reactor has been shut down for months?

Exercise:

Problem:

How can a nuclear reactor contain many critical masses and not go supercritical? What methods are used to control the fission in the reactor?

Exercise:

Problem:

Why can heavy nuclei with odd numbers of neutrons be induced to fission with thermal neutrons, whereas those with even numbers of neutrons require more energy input to induce fission?

Exercise:

Problem:

Why is a conventional fission nuclear reactor not able to explode as a bomb?

Problem Exercises

Exercise:

Problem:

(a) Calculate the energy released in the neutron-induced fission (similar to the spontaneous fission in [link])

Equation:

$$n + {}^{238}\text{U} \rightarrow {}^{96}\text{Sr} + {}^{140}\text{Xe} + 3n,$$

given $m({}^{96}\text{Sr}) = 95.921750\text{ u}$ and $m({}^{140}\text{Xe}) = 139.92164$. (b) This result is about 6 MeV greater than the result for spontaneous fission. Why? (c) Confirm that the total number of nucleons and total charge are conserved in this reaction.

Solution:

(a) 177.1 MeV

(b) Because the gain of an external neutron yields about 6 MeV, which is the average BE/A for heavy nuclei.

(c)
$A = 1 + 238 = 96 + 140 + 1 + 1 + 1,\ Z = 92 = 38 + 53,\ \text{efn} = 0 = 0$

Exercise:

Problem:

(a) Calculate the energy released in the neutron-induced fission reaction
Equation:

$$n + {}^{235}\text{U} \rightarrow {}^{92}\text{Kr} + {}^{142}\text{Ba} + 2n,$$

given $m({}^{92}\text{Kr}) = 91.926269\ \text{u}$ and $m({}^{142}\text{Ba}) = 141.916361\ \text{u}.$

(b) Confirm that the total number of nucleons and total charge are conserved in this reaction.

Exercise:

Problem:

(a) Calculate the energy released in the neutron-induced fission reaction
Equation:

$$n + {}^{239}\text{Pu} \rightarrow {}^{96}\text{Sr} + {}^{140}\text{Ba} + 4n,$$

given $m({}^{96}\text{Sr}) = 95.921750\ \text{u}$ and $m({}^{140}\text{Ba}) = 139.910581\ \text{u}.$

(b) Confirm that the total number of nucleons and total charge are conserved in this reaction.

Solution:

(a) 180.6 MeV

(b)
$A = 1 + 239 = 96 + 140 + 1 + 1 + 1 + 1,\ Z = 94 = 38 + 56,\ \text{efn} = 0 = 0$

Exercise:

Problem:

Confirm that each of the reactions listed for plutonium breeding just following [link] conserves the total number of nucleons, the total charge, and electron family number.

Exercise:

Problem:

Breeding plutonium produces energy even before any plutonium is fissioned. (The primary purpose of the four nuclear reactors at Chernobyl was breeding plutonium for weapons. Electrical power was a by-product used by the civilian population.) Calculate the energy produced in each of the reactions listed for plutonium breeding just following [link]. The pertinent masses are $m(^{239}\text{U}) = 239.054289\text{ u}$, $m(^{239}\text{Np}) = 239.052932\text{ u}$, and $m(^{239}\text{Pu}) = 239.052157\text{ u}$.

Solution:

$^{238}\text{U} + n \ \rightarrow {}^{239}\text{U} + \gamma$ 4.81 MeV

$^{239}\text{U} \rightarrow {}^{239}\text{Np} + \beta^{-} + v_{e}$ 0.753 MeV

$^{239}\text{Np} \rightarrow {}^{239}\text{Pu} + \beta^{-} + v_{e}$ 0.211 MeV

Exercise:

Problem:

The naturally occurring radioactive isotope ^{232}Th does not make good fission fuel, because it has an even number of neutrons; however, it can be bred into a suitable fuel (much as ^{238}U is bred into ^{239}P).

(a) What are Z and N for ^{232}Th?

(b) Write the reaction equation for neutron captured by ^{232}Th and identify the nuclide ^{A}X produced in $n + {}^{232}\text{Th} \rightarrow {}^{A}X + \gamma$.

(c) The product nucleus β^{-} decays, as does its daughter. Write the decay equations for each, and identify the final nucleus.

(d) Confirm that the final nucleus has an odd number of neutrons, making it a better fission fuel.

(e) Look up the half-life of the final nucleus to see if it lives long enough to be a useful fuel.

Exercise:

Problem:

The electrical power output of a large nuclear reactor facility is 900 MW. It has a 35.0% efficiency in converting nuclear power to electrical.

(a) What is the thermal nuclear power output in megawatts?

(b) How many ^{235}U nuclei fission each second, assuming the average fission produces 200 MeV?

(c) What mass of ^{235}U is fissioned in one year of full-power operation?

Solution:

(a) $2.57 \times 10^3 \text{ MW}$

(b) $8.03 \times 10^{19} \text{ fission/s}$

(c) 991 kg

Exercise:

Problem:

A large power reactor that has been in operation for some months is turned off, but residual activity in the core still produces 150 MW of power. If the average energy per decay of the fission products is 1.00 MeV, what is the core activity in curies?

Glossary

breeder reactors
: reactors that are designed specifically to make plutonium

breeding
: reaction process that produces ^{239}Pu

criticality
: condition in which a chain reaction easily becomes self-sustaining

critical mass

minimum amount necessary for self-sustained fission of a given nuclide

fission fragments
a daughter nuclei

liquid drop model
a model of nucleus (only to understand some of its features) in which nucleons in a nucleus act like atoms in a drop

nuclear fission
reaction in which a nucleus splits

neutron-induced fission
fission that is initiated after the absorption of neutron

supercriticality
an exponential increase in fissions

Nuclear Weapons

- Discuss different types of fission and thermonuclear bombs.
- Explain the ill effects of nuclear explosion.

The world was in turmoil when fission was discovered in 1938. The discovery of fission, made by two German physicists, Otto Hahn and Fritz Strassman, was quickly verified by two Jewish refugees from Nazi Germany, Lise Meitner and her nephew Otto Frisch. Fermi, among others, soon found that not only did neutrons induce fission; more neutrons were produced during fission. The possibility of a self-sustained chain reaction was immediately recognized by leading scientists the world over. The enormous energy known to be in nuclei, but considered inaccessible, now seemed to be available on a large scale.

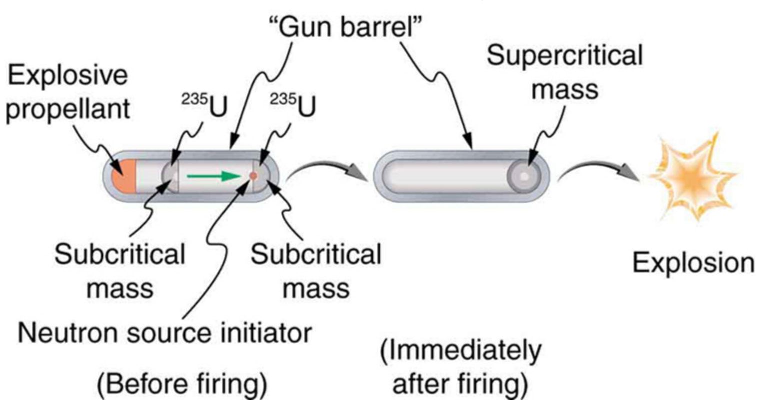
Within months after the announcement of the discovery of fission, Adolf Hitler banned the export of uranium from newly occupied Czechoslovakia. It seemed that the military value of uranium had been recognized in Nazi Germany, and that a serious effort to build a nuclear bomb had begun.

Alarmed scientists, many of them who fled Nazi Germany, decided to take action. None was more famous or revered than Einstein. It was felt that his help was needed to get the American government to make a serious effort at nuclear weapons as a matter of survival. Leo Szilard, an escaped Hungarian physicist, took a draft of a letter to Einstein, who, although pacifistic, signed the final version. The letter was for President Franklin Roosevelt, warning of the German potential to build extremely powerful bombs of a new type. It was sent in August of 1939, just before the German invasion of Poland that marked the start of World War II.

It was not until December 6, 1941, the day before the Japanese attack on Pearl Harbor, that the United States made a massive commitment to building a nuclear bomb. The top secret Manhattan Project was a crash program aimed at beating the Germans. It was carried out in remote locations, such as Los Alamos, New Mexico, whenever possible, and eventually came to cost billions of dollars and employ the efforts of more than 100,000 people. J. Robert Oppenheimer (1904–1967), whose talent and ambitions made him ideal, was chosen to head the project. The first

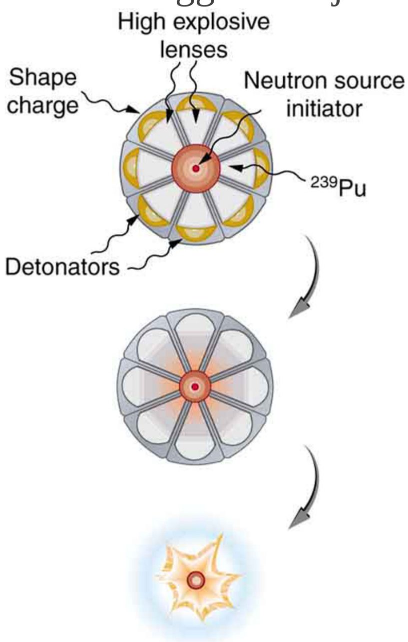
major step was made by Enrico Fermi and his group in December 1942, when they achieved the first self-sustained nuclear reactor. This first “atomic pile”, built in a squash court at the University of Chicago, used carbon blocks to thermalize neutrons. It not only proved that the chain reaction was possible, it began the era of nuclear reactors. Glenn Seaborg, an American chemist and physicist, received the Nobel Prize in physics in 1951 for discovery of several transuranic elements, including plutonium. Carbon-moderated reactors are relatively inexpensive and simple in design and are still used for breeding plutonium, such as at Chernobyl, where two such reactors remain in operation.

Plutonium was recognized as easier to fission with neutrons and, hence, a superior fission material very early in the Manhattan Project. Plutonium availability was uncertain, and so a uranium bomb was developed simultaneously. [link] shows a gun-type bomb, which takes two subcritical uranium masses and blows them together. To get an appreciable yield, the critical mass must be held together by the explosive charges inside the cannon barrel for a few microseconds. Since the buildup of the uranium chain reaction is relatively slow, the device to hold the critical mass together can be relatively simple. Owing to the fact that the rate of spontaneous fission is low, a neutron source is triggered at the same time the critical mass is assembled.



A gun-type fission bomb for ^{235}U utilizes two subcritical masses forced together by explosive charges inside a cannon barrel. The energy yield depends on the amount of uranium and the time it can be held together before it disassembles itself.

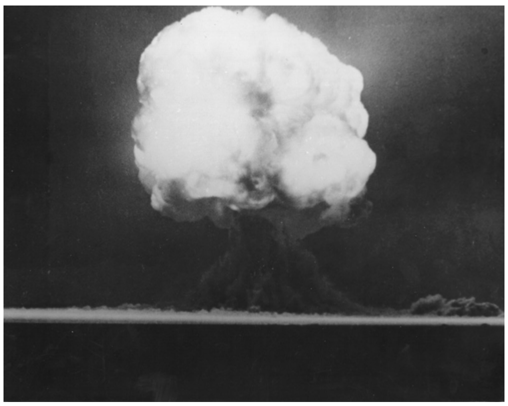
Plutonium’s special properties necessitated a more sophisticated critical mass assembly, shown schematically in [link]. A spherical mass of plutonium is surrounded by shape charges (high explosives that release most of their blast in one direction) that implode the plutonium, crushing it into a smaller volume to form a critical mass. The implosion technique is faster and more effective, because it compresses three-dimensionally rather than one-dimensionally as in the gun-type bomb. Again, a neutron source must be triggered at just the correct time to initiate the chain reaction.



An implosion created by high explosives compresses a sphere of ^{239}Pu into a critical mass. The superior fissionability of plutonium has made it the

universal bomb material.

Owing to its complexity, the plutonium bomb needed to be tested before there could be any attempt to use it. On July 16, 1945, the test named Trinity was conducted in the isolated Alamogordo Desert about 200 miles south of Los Alamos (see [link]). A new age had begun. The yield of this device was about 10 kilotons (kT), the equivalent of 5000 of the largest conventional bombs.



Trinity test (1945), the first nuclear bomb (credit: United States Department of Energy)

Although Germany surrendered on May 7, 1945, Japan had been steadfastly refusing to surrender for many months, forcing large casualties. Invasion plans by the Allies estimated a million casualties of their own and untold losses of Japanese lives. The bomb was viewed as a way to end the war. The first was a uranium bomb dropped on Hiroshima on August 6. Its yield of about 15 kT destroyed the city and killed an estimated 80,000 people, with 100,000 more being seriously injured (see [link]). The second was a plutonium bomb dropped on Nagasaki only three days later, on August 9. Its 20 kT yield killed at least 50,000 people, something less than Hiroshima because of the hilly terrain and the fact that it was a few kilometers off target. The Japanese were told that one bomb a week would be dropped

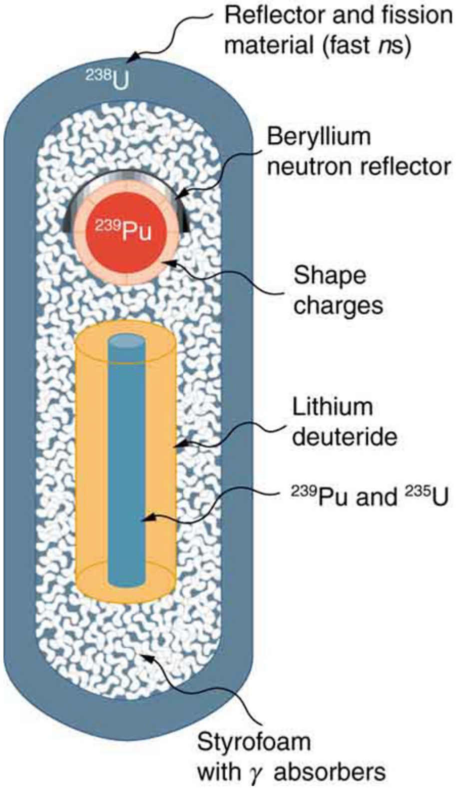
until they surrendered unconditionally, which they did on August 14. In actuality, the United States had only enough plutonium for one more and as yet unassembled bomb.

Destruction in Hiroshima (credit: United States Federal Government)

Knowing that fusion produces several times more energy per kilogram of fuel than fission, some scientists pushed the idea of a fusion bomb starting very early on. Calling this bomb the Super, they realized that it could have another advantage over fission—high-energy neutrons would aid fusion, while they are ineffective in ^{239}Pu fission. Thus the fusion bomb could be virtually unlimited in energy release. The first such bomb was detonated by the United States on October 31, 1952, at Eniwetok Atoll with a yield of 10 megatons (MT), about 670 times that of the fission bomb that destroyed Hiroshima. The Soviets followed with a fusion device of their own in August 1953, and a weapons race, beyond the aim of this text to discuss, continued until the end of the Cold War.

[link] shows a simple diagram of how a thermonuclear bomb is constructed. A fission bomb is exploded next to fusion fuel in the solid form of lithium deuteride. Before the shock wave blows it apart, γ rays heat and compress the fuel, and neutrons create tritium through the reaction $n+^6\text{Li}\to^3\text{H}+^4\text{He}$. Additional fusion and fission fuels are enclosed in a dense shell of ^{238}U. The shell reflects some of the neutrons back into the fuel to enhance its fusion, but at high internal temperatures fast neutrons are

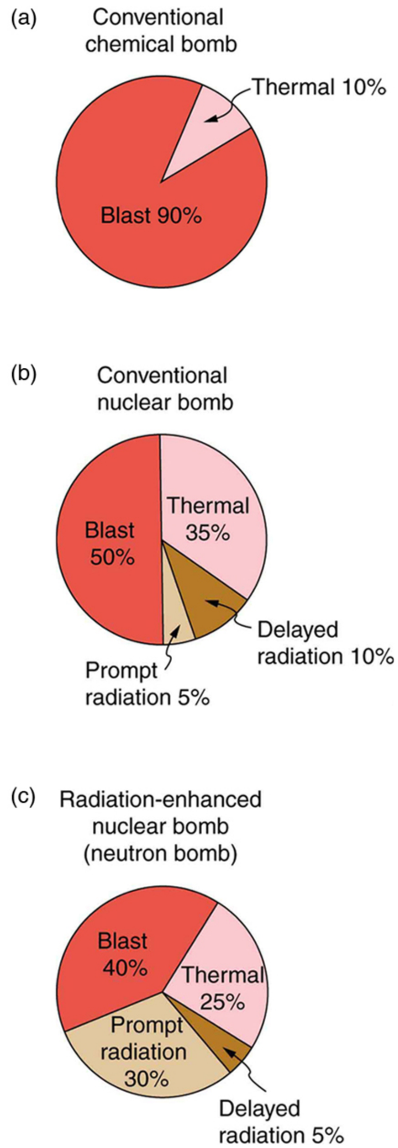
created that also cause the plentiful and inexpensive ^{238}U to fission, part of what allows thermonuclear bombs to be so large.



This schematic of a fusion bomb (H-bomb) gives some idea of how the ^{239}Pu fission trigger is used to ignite fusion fuel. Neutrons and γ rays transmit energy to the fusion fuel, create tritium from deuterium, and heat and compress the fusion fuel. The outer shell of ^{238}U serves to reflect some neutrons back into the fuel, causing more fusion,

and it boosts the energy output by fissioning itself when neutron energies become high enough.

The energy yield and the types of energy produced by nuclear bombs can be varied. Energy yields in current arsenals range from about 0.1 kT to 20 MT, although the Soviets once detonated a 67 MT device. Nuclear bombs differ from conventional explosives in more than size. [link] shows the approximate fraction of energy output in various forms for conventional explosives and for two types of nuclear bombs. Nuclear bombs put a much larger fraction of their output into thermal energy than do conventional bombs, which tend to concentrate the energy in blast. Another difference is the immediate and residual radiation energy from nuclear weapons. This can be adjusted to put more energy into radiation (the so-called neutron bomb) so that the bomb can be used to irradiate advancing troops without killing friendly troops with blast and heat.



Approximate fractions of energy output by conventional and two types of nuclear weapons. In addition to yielding more energy than conventional weapons, nuclear

bombs put a much larger fraction into thermal energy. This can be adjusted to enhance the radiation output to be more effective against troops. An enhanced radiation bomb is also called a neutron bomb.

At its peak in 1986, the combined arsenals of the United States and the Soviet Union totaled about 60,000 nuclear warheads. In addition, the British, French, and Chinese each have several hundred bombs of various sizes, and a few other countries have a small number. Nuclear weapons are generally divided into two categories. Strategic nuclear weapons are those intended for military targets, such as bases and missile complexes, and moderate to large cities. There were about 20,000 strategic weapons in 1988. Tactical weapons are intended for use in smaller battles. Since the collapse of the Soviet Union and the end of the Cold War in 1989, most of the 32,000 tactical weapons (including Cruise missiles, artillery shells, land mines, torpedoes, depth charges, and backpacks) have been demobilized, and parts of the strategic weapon systems are being dismantled with warheads and missiles being disassembled. According to the Treaty of Moscow of 2002, Russia and the United States have been required to reduce their strategic nuclear arsenal down to about 2000 warheads each.

A few small countries have built or are capable of building nuclear bombs, as are some terrorist groups. Two things are needed—a minimum level of technical expertise and sufficient fissionable material. The first is easy. Fissionable material is controlled but is also available. There are international agreements and organizations that attempt to control nuclear proliferation, but it is increasingly difficult given the availability of

fissionable material and the small amount needed for a crude bomb. The production of fissionable fuel itself is technologically difficult. However, the presence of large amounts of such material worldwide, though in the hands of a few, makes control and accountability crucial.

Section Summary

- There are two types of nuclear weapons—fission bombs use fission alone, whereas thermonuclear bombs use fission to ignite fusion.
- Both types of weapons produce huge numbers of nuclear reactions in a very short time.
- Energy yields are measured in kilotons or megatons of equivalent conventional explosives and range from 0.1 kT to more than 20 MT.
- Nuclear bombs are characterized by far more thermal output and nuclear radiation output than conventional explosives.

Conceptual Questions

Exercise:

Problem:

What are some of the reasons that plutonium rather than uranium is used in all fission bombs and as the trigger in all fusion bombs?

Exercise:

Problem:

Use the laws of conservation of momentum and energy to explain how a shape charge can direct most of the energy released in an explosion in a specific direction. (Note that this is similar to the situation in guns and cannons—most of the energy goes into the bullet.)

Exercise:

Problem:

How does the lithium deuteride in the thermonuclear bomb shown in [link] supply tritium (^3H) as well as deuterium (^2H)?

Exercise:

Problem:

Fallout from nuclear weapons tests in the atmosphere is mainly ^{90}Sr and ^{137}Cs, which have 28.6- and 32.2-y half-lives, respectively. Atmospheric tests were terminated in most countries in 1963, although China only did so in 1980. It has been found that environmental activities of these two isotopes are decreasing faster than their half-lives. Why might this be?

Problems & Exercises

Exercise:

Problem: Find the mass converted into energy by a 12.0-kT bomb.

Solution:

0.56 g

Exercise:

Problem: What mass is converted into energy by a 1.00-MT bomb?

Exercise:

Problem:

Fusion bombs use neutrons from their fission trigger to create tritium fuel in the reaction $n +^6 \text{Li} \to^3 \text{H} +^4 \text{He}$. What is the energy released by this reaction in MeV?

Solution:

4.781 MeV

Exercise:

Problem:

It is estimated that the total explosive yield of all the nuclear bombs in existence currently is about 4,000 MT.

(a) Convert this amount of energy to kilowatt-hours, noting that $1\ \text{kW} \cdot \text{h} = 3.60 \times 10^6\ \text{J}$.

(b) What would the monetary value of this energy be if it could be converted to electricity costing 10 cents per kW·h?

Exercise:

Problem:

A radiation-enhanced nuclear weapon (or neutron bomb) can have a smaller total yield and still produce more prompt radiation than a conventional nuclear bomb. This allows the use of neutron bombs to kill nearby advancing enemy forces with radiation without blowing up your own forces with the blast. For a 0.500-kT radiation-enhanced weapon and a 1.00-kT conventional nuclear bomb: (a) Compare the blast yields. (b) Compare the prompt radiation yields.

Solution:

(a) Blast yields $2.1 \times 10^{12}\ \text{J}$ to $8.4 \times 10^{11}\ \text{J}$, or 2.5 to 1, conventional to radiation enhanced.

(b) Prompt radiation yields $6.3 \times 10^{11}\ \text{J}$ to $2.1 \times 10^{11}\ \text{J}$, or 3 to 1, radiation enhanced to conventional.

Exercise:

Problem:

(a) How many ^{239}Pu nuclei must fission to produce a 20.0-kT yield, assuming 200 MeV per fission? (b) What is the mass of this much ^{239}Pu?

Exercise:

Problem:

Assume one-fourth of the yield of a typical 320-kT strategic bomb comes from fission reactions averaging 200 MeV and the remainder from fusion reactions averaging 20 MeV.

(a) Calculate the number of fissions and the approximate mass of uranium and plutonium fissioned, taking the average atomic mass to be 238.

(b) Find the number of fusions and calculate the approximate mass of fusion fuel, assuming an average total atomic mass of the two nuclei in each reaction to be 5.

(c) Considering the masses found, does it seem reasonable that some missiles could carry 10 warheads? Discuss, noting that the nuclear fuel is only a part of the mass of a warhead.

Solution:

(a) 1.1×10^{25} fissions , 4.4 kg

(b) 3.2×10^{26} fusions , 2.7 kg

(c) The nuclear fuel totals only 6 kg, so it is quite reasonable that some missiles carry 10 overheads. The mass of the fuel would only be 60 kg and therefore the mass of the 10 warheads, weighing about 10 times the nuclear fuel, would be only 1500 lbs. If the fuel for the missiles weighs 5 times the total weight of the warheads, the missile would weigh about 9000 lbs or 4.5 tons. This is not an unreasonable weight for a missile.

Exercise:

Problem:

This problem gives some idea of the magnitude of the energy yield of a small tactical bomb. Assume that half the energy of a 1.00-kT nuclear depth charge set off under an aircraft carrier goes into lifting it out of the water—that is, into gravitational potential energy. How high is the carrier lifted if its mass is 90,000 tons?

Exercise:

Problem:

It is estimated that weapons tests in the atmosphere have deposited approximately 9 MCi of $^{90}\mathrm{Sr}$ on the surface of the earth. Find the mass of this amount of $^{90}\mathrm{Sr}$.

Solution:

$7 \times 10^4 \text{ g}$

Exercise:

Problem:

A 1.00-MT bomb exploded a few kilometers above the ground deposits 25.0% of its energy into radiant heat.

(a) Find the calories per cm^2 at a distance of 10.0 km by assuming a uniform distribution over a spherical surface of that radius.

(b) If this heat falls on a person's body, what temperature increase does it cause in the affected tissue, assuming it is absorbed in a layer 1.00-cm deep?

Exercise:

Problem: Integrated Concepts

One scheme to put nuclear weapons to nonmilitary use is to explode them underground in a geologically stable region and extract the

geothermal energy for electricity production. There was a total yield of about 4,000 MT in the combined arsenals in 2006. If 1.00 MT per day could be converted to electricity with an efficiency of 10.0%:

(a) What would the average electrical power output be?

(b) How many years would the arsenal last at this rate?

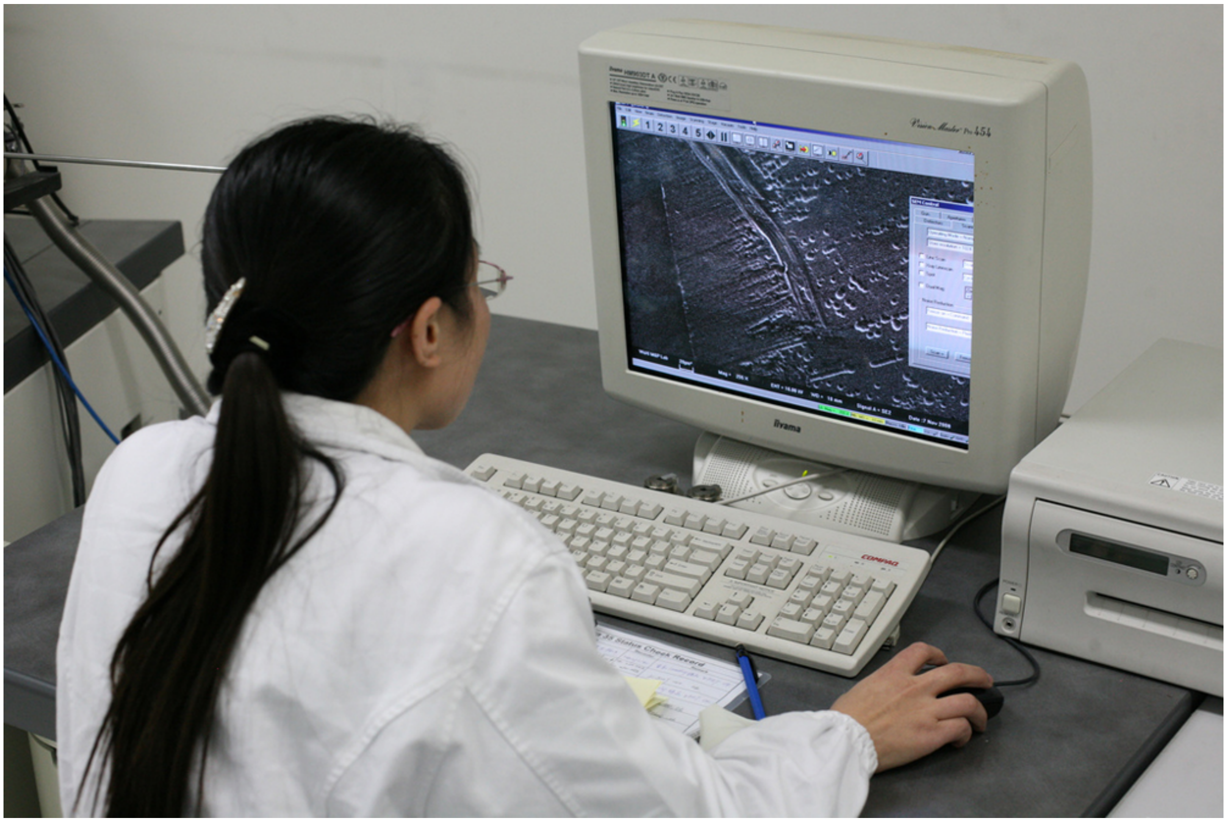
Solution:

(a) 4.86×10^9 W

(b) 11.0 y

Introduction to Vision and Optical Instruments
class="introduction"

A scientist examines minute details on the surface of a disk drive at a magnification of 100,000 times. The image was produced using an electron microscope. (credit: Robert Scoble)



Explore how the image on the computer screen is formed. How is the image formation on the computer screen different from the image formation in your eye as you look down the microscope? How can videos of living cell processes be taken for viewing later on, and by many different people?

Seeing faces and objects we love and cherish is a delight—one's favorite teddy bear, a picture on the wall, or the sun rising over the mountains. Intricate images help us understand nature and are invaluable for developing techniques and technologies in order to improve the quality of life. The image of a red blood cell that almost fills the cross-sectional area of a tiny capillary makes us wonder how blood makes it through and not get stuck. We are able to see bacteria and viruses and understand their structure. It is the knowledge of physics that provides fundamental understanding and models required to develop new techniques and instruments. Therefore, physics is called an *enabling science*—a science that enables development and advancement in other areas. It is through optics and imaging that physics enables advancement in major areas of biosciences. This chapter illustrates the enabling nature of physics through an understanding of how a

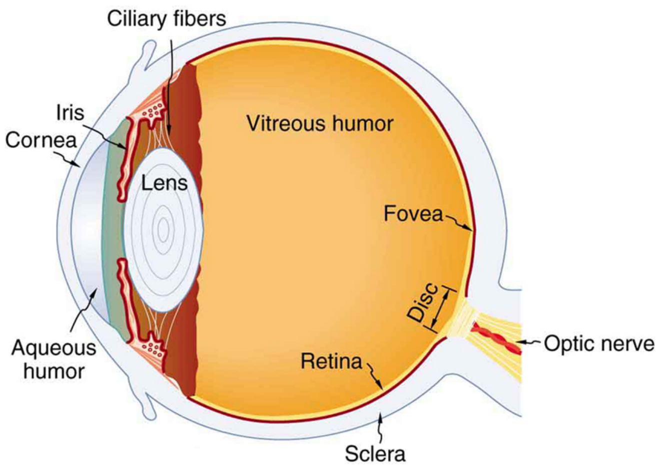
human eye is able to see and how we are able to use optical instruments to see beyond what is possible with the naked eye. It is convenient to categorize these instruments on the basis of geometric optics (see Geometric Optics) and wave optics (see Wave Optics).

Physics of the Eye

- Explain the image formation by the eye.
- Explain why peripheral images lack detail and color.
- Define refractive indices.
- Analyze the accommodation of the eye for distant and near vision.

The eye is perhaps the most interesting of all optical instruments. The eye is remarkable in how it forms images and in the richness of detail and color it can detect. However, our eyes commonly need some correction, to reach what is called “normal” vision, but should be called ideal rather than normal. Image formation by our eyes and common vision correction are easy to analyze with the optics discussed in Geometric Optics.

[link] shows the basic anatomy of the eye. The cornea and lens form a system that, to a good approximation, acts as a single thin lens. For clear vision, a real image must be projected onto the light-sensitive retina, which lies at a fixed distance from the lens. The lens of the eye adjusts its power to produce an image on the retina for objects at different distances. The center of the image falls on the fovea, which has the greatest density of light receptors and the greatest acuity (sharpness) in the visual field. The variable opening (or pupil) of the eye along with chemical adaptation allows the eye to detect light intensities from the lowest observable to 10^{10} times greater (without damage). This is an incredible range of detection. Our eyes perform a vast number of functions, such as sense direction, movement, sophisticated colors, and distance. Processing of visual nerve impulses begins with interconnections in the retina and continues in the brain. The optic nerve conveys signals received by the eye to the brain.



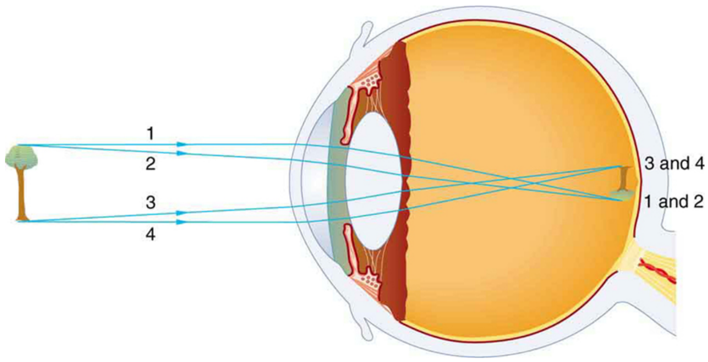
The cornea and lens of an eye act together to form a real image on the light-sensing retina, which has its densest concentration of receptors in the fovea and a blind spot over the optic nerve. The power of the lens of an eye is adjustable to provide an image on the retina for varying object distances. Layers of tissues with varying indices of refraction in the lens are shown here. However, they have been omitted from other pictures for clarity.

Refractive indices are crucial to image formation using lenses. [link] shows refractive indices relevant to the eye. The biggest change in the refractive index, and bending of rays, occurs at the cornea rather than the lens. The ray diagram in [link] shows image formation by the cornea and lens of the eye. The rays bend according to the refractive indices provided in [link]. The cornea provides about two-thirds of the power of the eye, owing to the fact that speed of light changes considerably while traveling from air into cornea. The lens provides the remaining power needed to produce an image on the retina. The cornea and lens can be treated as a single thin lens, even

though the light rays pass through several layers of material (such as cornea, aqueous humor, several layers in the lens, and vitreous humor), changing direction at each interface. The image formed is much like the one produced by a single convex lens. This is a case 1 image. Images formed in the eye are inverted but the brain inverts them once more to make them seem upright.

Material	**Index of Refraction**
Water	1.33
Air	1.0
Cornea	1.38
Aqueous humor	1.34
Lens	1.41 average (varies throughout the lens, greatest in center)
Vitreous humor	1.34

Refractive Indices Relevant to the Eye

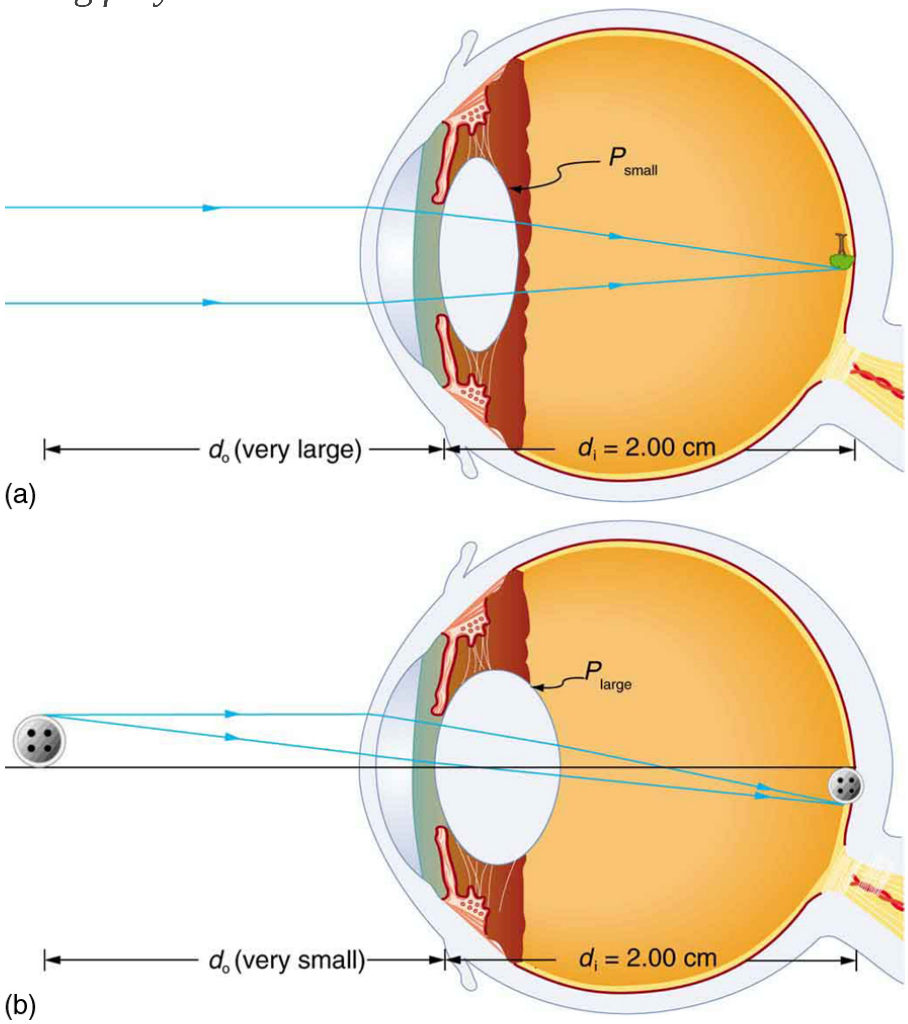


An image is formed on the retina with light rays converging most at the cornea and upon entering and exiting the lens. Rays from the top and bottom of the object are traced and produce an inverted real image on the retina. The distance to the object is drawn smaller than scale.

As noted, the image must fall precisely on the retina to produce clear vision — that is, the image distance d_i must equal the lens-to-retina distance. Because the lens-to-retina distance does not change, the image distance d_i must be the same for objects at all distances. The eye manages this by varying the power (and focal length) of the lens to accommodate for objects at various distances. The process of adjusting the eye's focal length is called **accommodation**. A person with normal (ideal) vision can see objects clearly at distances ranging from 25 cm to essentially infinity. However, although the near point (the shortest distance at which a sharp focus can be obtained) increases with age (becoming meters for some older people), we will consider it to be 25 cm in our treatment here.

[link] shows the accommodation of the eye for distant and near vision. Since light rays from a nearby object can diverge and still enter the eye, the lens must be more converging (more powerful) for close vision than for distant vision. To be more converging, the lens is made thicker by the action of the ciliary muscle surrounding it. The eye is most relaxed when viewing

distant objects, one reason that microscopes and telescopes are designed to produce distant images. Vision of very distant objects is called *totally relaxed*, while close vision is termed *accommodated*, with the closest vision being *fully accommodated*.



Relaxed and accommodated vision for distant and close objects. (a) Light rays from the same point on a distant object must be nearly parallel while entering the eye and more easily converge to produce an image on the retina. (b) Light rays from a nearby object can diverge more and still enter the eye. A more powerful lens is needed to converge them on the retina than if they were parallel.

We will use the thin lens equations to examine image formation by the eye quantitatively. First, note the power of a lens is given as $p = 1/f$, so we rewrite the thin lens equations as

Equation:

$$P = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}}$$

and

Equation:

$$\frac{h_{\text{i}}}{h_{\text{o}}} = -\frac{d_{\text{i}}}{d_{\text{o}}} = m.$$

We understand that d_{i} must equal the lens-to-retina distance to obtain clear vision, and that normal vision is possible for objects at distances $d_{\text{o}} = 25\text{ cm}$ to infinity.

Note:

Take-Home Experiment: The Pupil

Look at the central transparent area of someone’s eye, the pupil, in normal room light. Estimate the diameter of the pupil. Now turn off the lights and darken the room. After a few minutes turn on the lights and promptly estimate the diameter of the pupil. What happens to the pupil as the eye adjusts to the room light? Explain your observations.

The eye can detect an impressive amount of detail, considering how small the image is on the retina. To get some idea of how small the image can be, consider the following example.

Example:

Size of Image on Retina

What is the size of the image on the retina of a 1.20×10^{-2} cm diameter human hair, held at arm's length (60.0 cm) away? Take the lens-to-retina distance to be 2.00 cm.

Strategy

We want to find the height of the image h_i, given the height of the object is $h_\mathrm{o} = 1.20 \times 10^{-2}$ cm. We also know that the object is 60.0 cm away, so that $d_\mathrm{o} = 60.0$ cm. For clear vision, the image distance must equal the lens-to-retina distance, and so $d_\mathrm{i} = 2.00$ cm . The equation $\frac{h_\mathrm{i}}{h_\mathrm{o}} = -\frac{d_\mathrm{i}}{d_\mathrm{o}} = m$ can be used to find h_i with the known information.

Solution

The only unknown variable in the equation $\frac{h_\mathrm{i}}{h_\mathrm{o}} = -\frac{d_\mathrm{i}}{d_\mathrm{o}} = m$ is h_i:

Equation:

$$\frac{h_\mathrm{i}}{h_\mathrm{o}} = -\frac{d_\mathrm{i}}{d_\mathrm{o}}.$$

Rearranging to isolate h_i yields

Equation:

$$h_\mathrm{i} = -h_\mathrm{o} \cdot \frac{d_\mathrm{i}}{d_\mathrm{o}}.$$

Substituting the known values gives

Equation:

$$\begin{aligned} h_\mathrm{i} &= -(1.20 \times 10^{-2} \text{ cm}) \frac{2.00 \text{ cm}}{60.0 \text{ cm}} \\ &= -4.00 \times 10^{-4} \text{ cm}. \end{aligned}$$

Discussion

This truly small image is not the smallest discernible—that is, the limit to visual acuity is even smaller than this. Limitations on visual acuity have to do with the wave properties of light and will be discussed in the next chapter. Some limitation is also due to the inherent anatomy of the eye and processing that occurs in our brain.

Example:
Power Range of the Eye
Calculate the power of the eye when viewing objects at the greatest and smallest distances possible with normal vision, assuming a lens-to-retina distance of 2.00 cm (a typical value).

Strategy

For clear vision, the image must be on the retina, and so $d_\mathrm{i} = 2.00\text{ cm}$ here. For distant vision, $d_\mathrm{o} \approx \infty$, and for close vision, $d_\mathrm{o} = 25.0\text{ cm}$, as discussed earlier. The equation $P = \frac{1}{d_\mathrm{o}} + \frac{1}{d_\mathrm{i}}$ as written just above, can be used directly to solve for P in both cases, since we know d_i and d_o. Power has units of diopters, where $1\text{ D} = 1/\text{m}$, and so we should express all distances in meters.

Solution

For distant vision,

Equation:

$$P = \frac{1}{d_\mathrm{o}} + \frac{1}{d_\mathrm{i}} = \frac{1}{\infty} + \frac{1}{0.0200\text{ m}}.$$

Since $1/\infty = 0$, this gives

Equation:

$$P = 0 + 50.0/\text{m} = 50.0\text{ D (distant vision)}.$$

Now, for close vision,

Equation:

$$\begin{aligned} P &= \tfrac{1}{d_\mathrm{o}} + \tfrac{1}{d_\mathrm{i}} = \tfrac{1}{0.250\text{ m}} + \tfrac{1}{0.0200\text{ m}} \\ &= \tfrac{4.00}{\text{m}} + \tfrac{50.0}{\text{m}} = 4.00\text{ D} + 50.0\text{ D} \\ &= 54.0\text{ D (close vision)}. \end{aligned}$$

Discussion

For an eye with this typical 2.00 cm lens-to-retina distance, the power of the eye ranges from 50.0 D (for distant totally relaxed vision) to 54.0 D (for close fully accommodated vision), which is an 8% increase. This increase in power for close vision is consistent with the preceding

discussion and the ray tracing in [link]. An 8% ability to accommodate is considered normal but is typical for people who are about 40 years old. Younger people have greater accommodation ability, whereas older people gradually lose the ability to accommodate. When an optometrist identifies accommodation as a problem in elder people, it is most likely due to stiffening of the lens. The lens of the eye changes with age in ways that tend to preserve the ability to see distant objects clearly but do not allow the eye to accommodate for close vision, a condition called **presbyopia** (literally, elder eye). To correct this vision defect, we place a converging, positive power lens in front of the eye, such as found in reading glasses. Commonly available reading glasses are rated by their power in diopters, typically ranging from 1.0 to 3.5 D.

Section Summary

- Image formation by the eye is adequately described by the thin lens equations:
  Equation:
  $$P = \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} \text{ and } \frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}} = m.$$
- The eye produces a real image on the retina by adjusting its focal length and power in a process called accommodation.
- For close vision, the eye is fully accommodated and has its greatest power, whereas for distant vision, it is totally relaxed and has its smallest power.
- The loss of the ability to accommodate with age is called presbyopia, which is corrected by the use of a converging lens to add power for close vision.

Conceptual Questions

Exercise:

Problem:

If the lens of a person’s eye is removed because of cataracts (as has been done since ancient times), why would you expect a spectacle lens of about 16 D to be prescribed?

Exercise:

Problem:

A cataract is cloudiness in the lens of the eye. Is light dispersed or diffused by it?

Exercise:

Problem:

When laser light is shone into a relaxed normal-vision eye to repair a tear by spot-welding the retina to the back of the eye, the rays entering the eye must be parallel. Why?

Exercise:

Problem:

How does the power of a dry contact lens compare with its power when resting on the tear layer of the eye? Explain.

Exercise:

Problem:

Why is your vision so blurry when you open your eyes while swimming under water? How does a face mask enable clear vision?

Problem Exercises

Unless otherwise stated, the lens-to-retina distance is 2.00 cm.
Exercise:

Problem:

What is the power of the eye when viewing an object 50.0 cm away?

Solution:

52.0 D

Exercise:

Problem:

Calculate the power of the eye when viewing an object 3.00 m away.

Exercise:

Problem:

(a) The print in many books averages 3.50 mm in height. How high is the image of the print on the retina when the book is held 30.0 cm from the eye?

(b) Compare the size of the print to the sizes of rods and cones in the fovea and discuss the possible details observable in the letters. (The eye-brain system can perform better because of interconnections and higher order image processing.)

Solution:

(a) -0.233 mm

(b) The size of the rods and the cones is smaller than the image height, so we can distinguish letters on a page.

Exercise:

Problem:

Suppose a certain person's visual acuity is such that he can see objects clearly that form an image $4.00\ \mu m$ high on his retina. What is the maximum distance at which he can read the 75.0 cm high letters on the side of an airplane?

Exercise:

Problem:

People who do very detailed work close up, such as jewellers, often can see objects clearly at much closer distance than the normal 25 cm.

(a) What is the power of the eyes of a woman who can see an object clearly at a distance of only 8.00 cm?

(b) What is the size of an image of a 1.00 mm object, such as lettering inside a ring, held at this distance?

(c) What would the size of the image be if the object were held at the normal 25.0 cm distance?

Solution:

(a) $+62.5\ \mathrm{D}$

(b) $-0.250\ \mathrm{mm}$

(c) $-0.0800\ \mathrm{mm}$

Glossary

accommodation
: the ability of the eye to adjust its focal length is known as accommodation

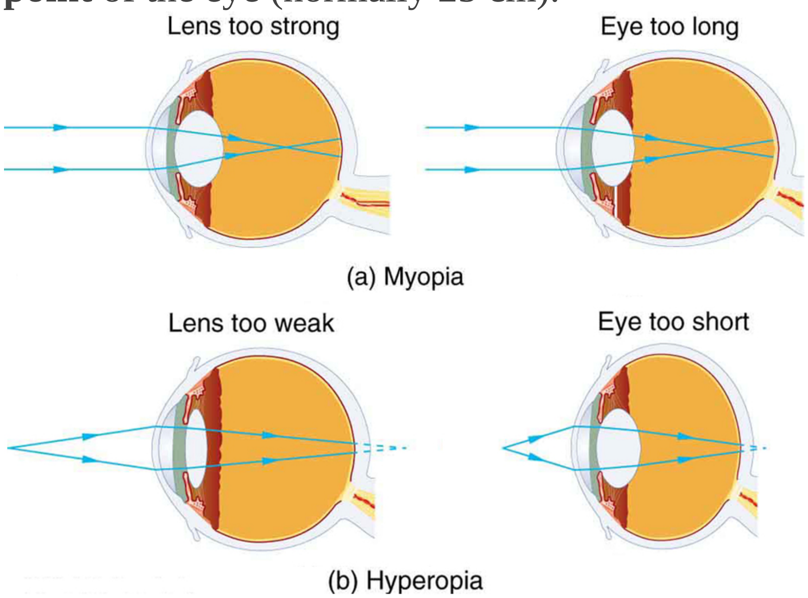
presbyopia

a condition in which the lens of the eye becomes progressively unable to focus on objects close to the viewer

Vision Correction

- Identify and discuss common vision defects.
- Explain nearsightedness and farsightedness corrections.
- Explain laser vision correction.

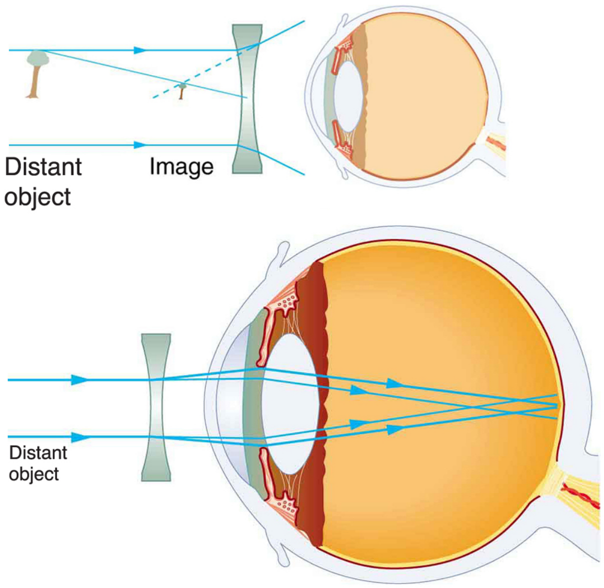
The need for some type of vision correction is very common. Common vision defects are easy to understand, and some are simple to correct. [link] illustrates two common vision defects. **Nearsightedness**, or **myopia**, is the inability to see distant objects clearly while close objects are clear. The eye overconverges the nearly parallel rays from a distant object, and the rays cross in front of the retina. More divergent rays from a close object are converged on the retina for a clear image. The distance to the farthest object that can be seen clearly is called the **far point** of the eye (normally infinity). **Farsightedness**, or **hyperopia**, is the inability to see close objects clearly while distant objects may be clear. A farsighted eye does not converge sufficient rays from a close object to make the rays meet on the retina. Less diverging rays from a distant object can be converged for a clear image. The distance to the closest object that can be seen clearly is called the **near point** of the eye (normally 25 cm).



(a) The nearsighted (myopic) eye converges rays from a distant object in front of the retina; thus, they are diverging when they

strike the retina, producing a blurry image. This can be caused by the lens of the eye being too powerful or the length of the eye being too great. (b) The farsighted (hyperopic) eye is unable to converge the rays from a close object by the time they strike the retina, producing blurry close vision. This can be caused by insufficient power in the lens or by the eye being too short.

Since the nearsighted eye over converges light rays, the correction for nearsightedness is to place a diverging spectacle lens in front of the eye. This reduces the power of an eye that is too powerful. Another way of thinking about this is that a diverging spectacle lens produces a case 3 image, which is closer to the eye than the object (see [link]). To determine the spectacle power needed for correction, you must know the person’s far point—that is, you must know the greatest distance at which the person can see clearly. Then the image produced by a spectacle lens must be at this distance or closer for the nearsighted person to be able to see it clearly. It is worth noting that wearing glasses does not change the eye in any way. The eyeglass lens is simply used to create an image of the object at a distance where the nearsighted person can see it clearly. Whereas someone not wearing glasses can see clearly *objects* that fall between their near point and their far point, someone wearing glasses can see *images* that fall between their near point and their far point.



Correction of nearsightedness requires a diverging lens that compensates for the overconvergence by the eye. The diverging lens produces an image closer to the eye than the object, so that the nearsighted person can see it clearly.

Example:
Correcting Nearsightedness

What power of spectacle lens is needed to correct the vision of a nearsighted person whose far point is 30.0 cm? Assume the spectacle (corrective) lens is held 1.50 cm away from the eye by eyeglass frames.

Strategy

You want this nearsighted person to be able to see very distant objects clearly. That means the spectacle lens must produce an image 30.0 cm from the eye for an object very far away. An image 30.0 cm from the eye will be 28.5 cm to the left of the spectacle lens (see [link]). Therefore, we

must get $d_i = -28.5\text{ cm}$ when $d_o \approx \infty$. The image distance is negative, because it is on the same side of the spectacle as the object.

Solution

Since d_i and d_o are known, the power of the spectacle lens can be found using $P = \frac{1}{d_o} + \frac{1}{d_i}$ as written earlier:

Equation:

$$P = \frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{\infty} + \frac{1}{-0.285\text{ m}}.$$

Since $1/\infty = 0$, we obtain:

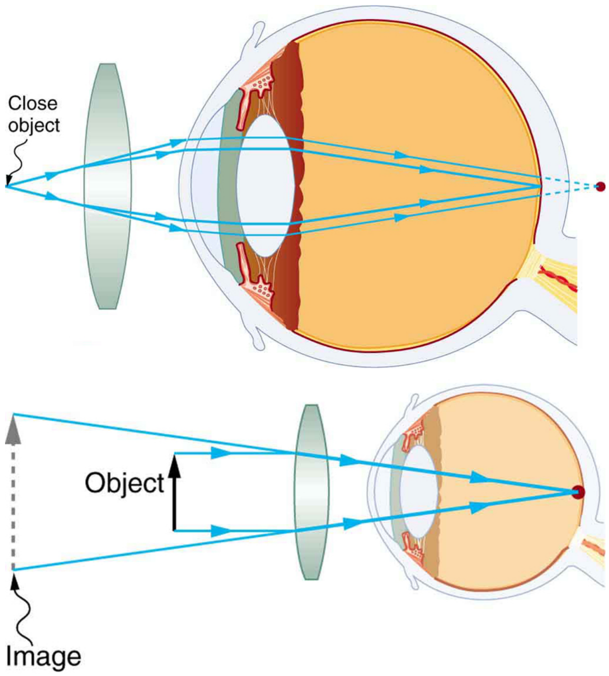
Equation:

$$P = 0 - 3.51/\text{m} = -3.51\text{ D}.$$

Discussion

The negative power indicates a diverging (or concave) lens, as expected. The spectacle produces a case 3 image closer to the eye, where the person can see it. If you examine eyeglasses for nearsighted people, you will find the lenses are thinnest in the center. Additionally, if you examine a prescription for eyeglasses for nearsighted people, you will find that the prescribed power is negative and given in units of diopters.

Since the farsighted eye under converges light rays, the correction for farsightedness is to place a converging spectacle lens in front of the eye. This increases the power of an eye that is too weak. Another way of thinking about this is that a converging spectacle lens produces a case 2 image, which is farther from the eye than the object (see [link]). To determine the spectacle power needed for correction, you must know the person’s near point—that is, you must know the smallest distance at which the person can see clearly. Then the image produced by a spectacle lens must be at this distance or farther for the farsighted person to be able to see it clearly.



Correction of farsightedness uses a converging lens that compensates for the under convergence by the eye. The converging lens produces an image farther from the eye than the object, so that the farsighted person can see it clearly.

Example:
Correcting Farsightedness

What power of spectacle lens is needed to allow a farsighted person, whose near point is 1.00 m, to see an object clearly that is 25.0 cm away? Assume the spectacle (corrective) lens is held 1.50 cm away from the eye by eyeglass frames.

Strategy

When an object is held 25.0 cm from the person's eyes, the spectacle lens must produce an image 1.00 m away (the near point). An image 1.00 m

from the eye will be 98.5 cm to the left of the spectacle lens because the spectacle lens is 1.50 cm from the eye (see [link]). Therefore, $d_\text{i} = -98.5 \text{ cm}$. The image distance is negative, because it is on the same side of the spectacle as the object. The object is 23.5 cm to the left of the spectacle, so that $d_\text{o} = 23.5 \text{ cm}$.

Solution

Since d_i and d_o are known, the power of the spectacle lens can be found using $P = \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}}$:

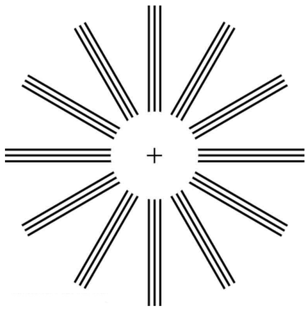
Equation:

$$\begin{array}{lll} P & = & \frac{1}{d_o} + \frac{1}{d_\text{i}} = \frac{1}{0.235 \text{ m}} + \frac{1}{-0.985 \text{ m}} \\ & = & 4.26 \text{ D} - 1.02 \text{ D} = 3.24 \text{ D}. \end{array}$$

Discussion

The positive power indicates a converging (convex) lens, as expected. The convex spectacle produces a case 2 image farther from the eye, where the person can see it. If you examine eyeglasses of farsighted people, you will find the lenses to be thickest in the center. In addition, a prescription of eyeglasses for farsighted people has a prescribed power that is positive.

Another common vision defect is **astigmatism**, an unevenness or asymmetry in the focus of the eye. For example, rays passing through a vertical region of the eye may focus closer than rays passing through a horizontal region, resulting in the image appearing elongated. This is mostly due to irregularities in the shape of the cornea but can also be due to lens irregularities or unevenness in the retina. Because of these irregularities, different parts of the lens system produce images at different locations. The eye-brain system can compensate for some of these irregularities, but they generally manifest themselves as less distinct vision or sharper images along certain axes. [link] shows a chart used to detect astigmatism. Astigmatism can be at least partially corrected with a spectacle having the opposite irregularity of the eye. If an eyeglass prescription has a cylindrical correction, it is there to correct astigmatism. The normal corrections for short- or farsightedness are spherical corrections, uniform along all axes.



This chart can detect astigmatism, unevenness in the focus of the eye. Check each of your eyes separately by looking at the center cross (without spectacles if you wear them). If lines along some axes appear darker or clearer than others, you have an astigmatism.

Contact lenses have advantages over glasses beyond their cosmetic aspects. One problem with glasses is that as the eye moves, it is not at a fixed distance from the spectacle lens. Contacts rest on and move with the eye, eliminating this problem. Because contacts cover a significant portion of the

cornea, they provide superior peripheral vision compared with eyeglasses. Contacts also correct some corneal astigmatism caused by surface irregularities. The tear layer between the smooth contact and the cornea fills in the irregularities. Since the index of refraction of the tear layer and the cornea are very similar, you now have a regular optical surface in place of an irregular one. If the curvature of a contact lens is not the same as the cornea (as may be necessary with some individuals to obtain a comfortable fit), the tear layer between the contact and cornea acts as a lens. If the tear layer is thinner in the center than at the edges, it has a negative power, for example. Skilled optometrists will adjust the power of the contact to compensate.

Laser vision correction has progressed rapidly in the last few years. It is the latest and by far the most successful in a series of procedures that correct vision by reshaping the cornea. As noted at the beginning of this section, the cornea accounts for about two-thirds of the power of the eye. Thus, small adjustments of its curvature have the same effect as putting a lens in front of the eye. To a reasonable approximation, the power of multiple lenses placed close together equals the sum of their powers. For example, a concave spectacle lens (for nearsightedness) having $P = -3.00\text{ D}$ has the same effect on vision as reducing the power of the eye itself by 3.00 D. So to correct the eye for nearsightedness, the cornea is flattened to reduce its power. Similarly, to correct for farsightedness, the curvature of the cornea is enhanced to increase the power of the eye—the same effect as the positive power spectacle lens used for farsightedness. Laser vision correction uses high intensity electromagnetic radiation to ablate (to remove material from the surface) and reshape the corneal surfaces.

Today, the most commonly used laser vision correction procedure is *Laser in situ Keratomileusis (LASIK)*. The top layer of the cornea is surgically peeled back and the underlying tissue ablated by multiple bursts of finely controlled ultraviolet radiation produced by an excimer laser. Lasers are used because they not only produce well-focused intense light, but they also emit very pure wavelength electromagnetic radiation that can be controlled more accurately than mixed wavelength light. The 193 nm wavelength UV commonly used is extremely and strongly absorbed by corneal tissue,

allowing precise evaporation of very thin layers. A computer controlled program applies more bursts, usually at a rate of 10 per second, to the areas that require deeper removal. Typically a spot less than 1 mm in diameter and about 0.3 μm in thickness is removed by each burst. Nearsightedness, farsightedness, and astigmatism can be corrected with an accuracy that produces normal distant vision in more than 90% of the patients, in many cases right away. The corneal flap is replaced; healing takes place rapidly and is nearly painless. More than 1 million Americans per year undergo LASIK (see [link]).

Laser vision correction is being performed using the LASIK procedure. Reshaping of the cornea by laser ablation is based on a careful assessment of the patient’s vision and is computer controlled. The

upper corneal layer is temporarily peeled back and minimally disturbed in LASIK, providing for more rapid and less painful healing of the less sensitive tissues below. (credit: U.S. Navy photo by Mass Communication Specialist 1st Class Brien Aho)

Section Summary

- Nearsightedness, or myopia, is the inability to see distant objects and is corrected with a diverging lens to reduce power.
- Farsightedness, or hyperopia, is the inability to see close objects and is corrected with a converging lens to increase power.
- In myopia and hyperopia, the corrective lenses produce images at a distance that the person can see clearly—the far point and near point, respectively.

Conceptual Questions

Exercise:

Problem:

It has become common to replace the cataract-clouded lens of the eye with an internal lens. This intraocular lens can be chosen so that the person has perfect distant vision. Will the person be able to read without glasses? If the person was nearsighted, is the power of the intraocular lens greater or less than the removed lens?

Exercise:

Problem:

If the cornea is to be reshaped (this can be done surgically or with contact lenses) to correct myopia, should its curvature be made greater or smaller? Explain. Also explain how hyperopia can be corrected.

Exercise:

Problem:

If there is a fixed percent uncertainty in LASIK reshaping of the cornea, why would you expect those people with the greatest correction to have a poorer chance of normal distant vision after the procedure?

Exercise:

Problem:

A person with presbyopia has lost some or all of the ability to accommodate the power of the eye. If such a person's distant vision is corrected with LASIK, will she still need reading glasses? Explain.

Problem Exercises

Exercise:

Problem:

What is the far point of a person whose eyes have a relaxed power of 50.5 D?

Solution:

2.00 m

Exercise:

Problem:

What is the near point of a person whose eyes have an accommodated power of 53.5 D?

Exercise:

Problem:

(a) A laser vision correction reshaping the cornea of a myopic patient reduces the power of his eye by 9.00 D, with a $\pm 5.0\%$ uncertainty in the final correction. What is the range of diopters for spectacle lenses that this person might need after LASIK procedure? (b) Was the person nearsighted or farsighted before the procedure? How do you know?

Solution:

(a) ± 0.45 D

(b) The person was nearsighted because the patient was myopic and the power was reduced.

Exercise:

Problem:

In a LASIK vision correction, the power of a patient's eye is increased by 3.00 D. Assuming this produces normal close vision, what was the patient's near point before the procedure?

Exercise:

Problem:

What was the previous far point of a patient who had laser vision correction that reduced the power of her eye by 7.00 D, producing normal distant vision for her?

Solution:

0.143 m

Exercise:

Problem:

A severely myopic patient has a far point of 5.00 cm. By how many diopters should the power of his eye be reduced in laser vision correction to obtain normal distant vision for him?

Exercise:

Problem:

A student's eyes, while reading the blackboard, have a power of 51.0 D. How far is the board from his eyes?

Solution:

1.00 m

Exercise:

Problem:

The power of a physician's eyes is 53.0 D while examining a patient. How far from her eyes is the feature being examined?

Exercise:

Problem:

A young woman with normal distant vision has a 10.0% ability to accommodate (that is, increase) the power of her eyes. What is the closest object she can see clearly?

Solution:

20.0 cm

Exercise:

Problem:

The far point of a myopic administrator is 50.0 cm. (a) What is the relaxed power of his eyes? (b) If he has the normal 8.00% ability to accommodate, what is the closest object he can see clearly?

Exercise:

Problem:

A very myopic man has a far point of 20.0 cm. What power contact lens (when on the eye) will correct his distant vision?

Solution:

–5.00 D

Exercise:

Problem:

Repeat the previous problem for eyeglasses held 1.50 cm from the eyes.

Exercise:

Problem:

A myopic person sees that her contact lens prescription is –4.00 D. What is her far point?

Solution:

25.0 cm

Exercise:

Problem:

Repeat the previous problem for glasses that are 1.75 cm from the eyes.

Exercise:

Problem:

The contact lens prescription for a mildly farsighted person is 0.750 D, and the person has a near point of 29.0 cm. What is the power of the tear layer between the cornea and the lens if the correction is ideal, taking the tear layer into account?

Solution:

–0.198 D

Exercise:

Problem:

A nearsighted man cannot see objects clearly beyond 20 cm from his eyes. How close must he stand to a mirror in order to see what he is doing when he shaves?

Exercise:

Problem:

A mother sees that her child's contact lens prescription is 0.750 D. What is the child's near point?

Solution:

30.8 cm

Exercise:

Problem:

Repeat the previous problem for glasses that are 2.20 cm from the eyes.

Exercise:

Problem:

The contact lens prescription for a nearsighted person is –4.00 D and the person has a far point of 22.5 cm. What is the power of the tear layer between the cornea and the lens if the correction is ideal, taking the tear layer into account?

Solution:

–0.444 D

Exercise:

Problem: **Unreasonable Results**

A boy has a near point of 50 cm and a far point of 500 cm. Will a –4.00 D lens correct his far point to infinity?

Glossary

nearsightedness
: another term for myopia, a visual defect in which distant objects appear blurred because their images are focused in front of the retina rather than being focused on the retina

myopia
: a visual defect in which distant objects appear blurred because their images are focused in front of the retina rather than being focused on the retina

far point

the object point imaged by the eye onto the retina in an unaccommodated eye

farsightedness

another term for hyperopia, the condition of an eye where incoming rays of light reach the retina before they converge into a focused image

hyperopia

the condition of an eye where incoming rays of light reach the retina before they converge into a focused image

near point

the point nearest the eye at which an object is accurately focused on the retina at full accommodation

astigmatism

the result of an inability of the cornea to properly focus an image onto the retina

laser vision correction

a medical procedure used to correct astigmatism and eyesight deficiencies such as myopia and hyperopia

Color and Color Vision

- Explain the simple theory of color vision.
- Outline the coloring properties of light sources.
- Describe the retinex theory of color vision.

The gift of vision is made richer by the existence of color. Objects and lights abound with thousands of hues that stimulate our eyes, brains, and emotions. Two basic questions are addressed in this brief treatment—what does color mean in scientific terms, and how do we, as humans, perceive it?

Simple Theory of Color Vision

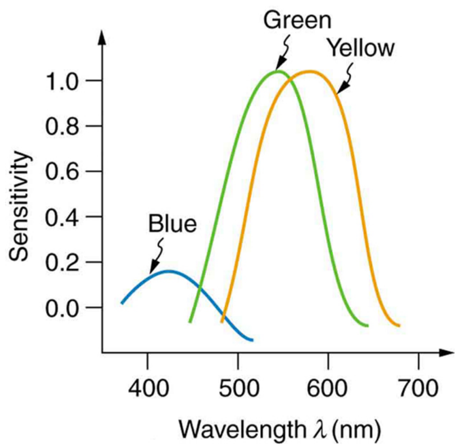
We have already noted that color is associated with the wavelength of visible electromagnetic radiation. When our eyes receive pure-wavelength light, we tend to see only a few colors. Six of these (most often listed) are red, orange, yellow, green, blue, and violet. These are the rainbow of colors produced when white light is dispersed according to different wavelengths. There are thousands of other **hues** that we can perceive. These include brown, teal, gold, pink, and white. One simple theory of color vision implies that all these hues are our eye's response to different combinations of wavelengths. This is true to an extent, but we find that color perception is even subtler than our eye's response for various wavelengths of light.

The two major types of light-sensing cells (photoreceptors) in the retina are **rods and cones**. Rods are more sensitive than cones by a factor of about 1000 and are solely responsible for peripheral vision as well as vision in very dark environments. They are also important for motion detection. There are about 120 million rods in the human retina. Rods do not yield color information. You may notice that you lose color vision when it is very dark, but you retain the ability to discern grey scales.

Note:
Take-Home Experiment: Rods and Cones

1. Go into a darkened room from a brightly lit room, or from outside in the Sun. How long did it take to start seeing shapes more clearly? What about color? Return to the bright room. Did it take a few minutes before you could see things clearly?
2. Demonstrate the sensitivity of foveal vision. Look at the letter G in the word ROGERS. What about the clarity of the letters on either side of G?

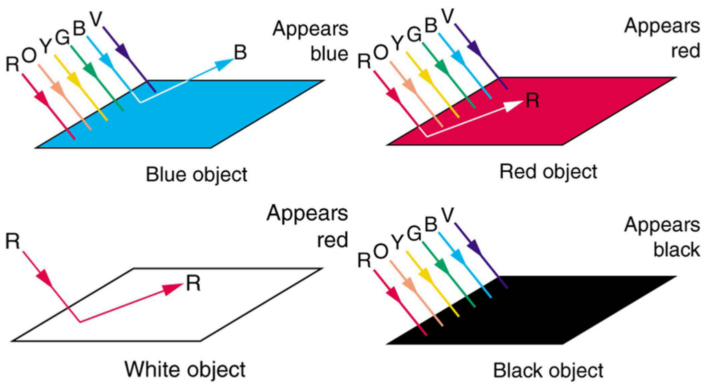
Cones are most concentrated in the fovea, the central region of the retina. There are no rods here. The fovea is at the center of the macula, a 5 mm diameter region responsible for our central vision. The cones work best in bright light and are responsible for high resolution vision. There are about 6 million cones in the human retina. There are three types of cones, and each type is sensitive to different ranges of wavelengths, as illustrated in [link]. A **simplified theory of color vision** is that there are three *primary colors* corresponding to the three types of cones. The thousands of other hues that we can distinguish among are created by various combinations of stimulations of the three types of cones. Color television uses a three-color system in which the screen is covered with equal numbers of red, green, and blue phosphor dots. The broad range of hues a viewer sees is produced by various combinations of these three colors. For example, you will perceive yellow when red and green are illuminated with the correct ratio of intensities. White may be sensed when all three are illuminated. Then, it would seem that all hues can be produced by adding three primary colors in various proportions. But there is an indication that color vision is more sophisticated. There is no unique set of three primary colors. Another set that works is yellow, green, and blue. A further indication of the need for a more complex theory of color vision is that various different combinations can produce the same hue. Yellow can be sensed with yellow light, or with a combination of red and green, and also with white light from which violet has been removed. The three-primary-colors aspect of color vision is well established; more sophisticated theories expand on it rather than deny it.



The image shows the relative sensitivity of the three types of cones, which are named according to wavelengths of greatest sensitivity. Rods are about 1000 times more sensitive, and their curve peaks at about 500 nm. Evidence for the three types of cones comes from direct measurements in animal and human eyes and testing of color blind people.

Consider why various objects display color—that is, why are feathers blue and red in a crimson rosella? The *true color of an object* is defined by its absorptive or reflective characteristics. [link] shows white light falling on three different objects, one pure blue, one pure red, and one black, as well as pure red light falling on a white object. Other hues are created by more

complex absorption characteristics. Pink, for example on a galah cockatoo, can be due to weak absorption of all colors except red. An object can appear a different color under non-white illumination. For example, a pure blue object illuminated with pure red light will *appear* black, because it absorbs all the red light falling on it. But, the true color of the object is blue, which is independent of illumination.



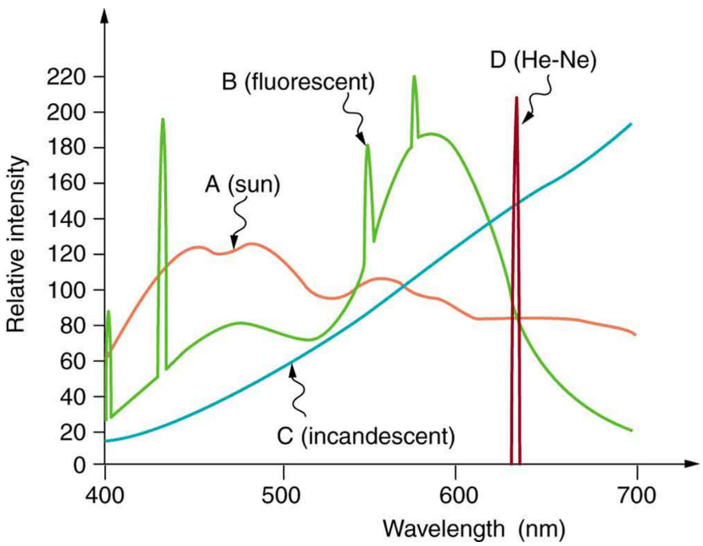
Absorption characteristics determine the true color of an object. Here, three objects are illuminated by white light, and one by pure red light. White is the equal mixture of all visible wavelengths; black is the absence of light.

Similarly, *light sources have colors* that are defined by the wavelengths they produce. A helium-neon laser emits pure red light. In fact, the phrase “pure red light” is defined by having a sharp constrained spectrum, a characteristic of laser light. The Sun produces a broad yellowish spectrum, fluorescent lights emit bluish-white light, and incandescent lights emit reddish-white hues as seen in [link]. As you would expect, you sense these colors when viewing the light source directly or when illuminating a white object with them. All of this fits neatly into the simplified theory that a combination of wavelengths produces various hues.

Note:

Take-Home Experiment: Exploring Color Addition

This activity is best done with plastic sheets of different colors as they allow more light to pass through to our eyes. However, thin sheets of paper and fabric can also be used. Overlay different colors of the material and hold them up to a white light. Using the theory described above, explain the colors you observe. You could also try mixing different crayon colors.

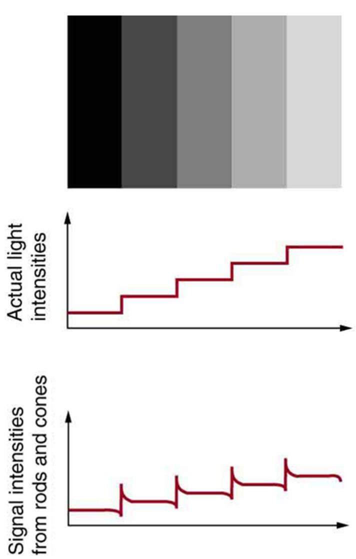


Emission spectra for various light sources are shown. Curve A is average sunlight at Earth's surface, curve B is light from a fluorescent lamp, and curve C is the output of an incandescent light. The spike for a helium-neon laser (curve D) is due to its pure wavelength emission. The spikes in the fluorescent output are due to atomic spectra—a topic that will be explored later.

Color Constancy and a Modified Theory of Color Vision

The eye-brain color-sensing system can, by comparing various objects in its view, perceive the true color of an object under varying lighting conditions—an ability that is called **color constancy**. We can sense that a white tablecloth, for example, is white whether it is illuminated by sunlight, fluorescent light, or candlelight. The wavelengths entering the eye are quite different in each case, as the graphs in [link] imply, but our color vision can detect the true color by comparing the tablecloth with its surroundings.

Theories that take color constancy into account are based on a large body of anatomical evidence as well as perceptual studies. There are nerve connections among the light receptors on the retina, and there are far fewer nerve connections to the brain than there are rods and cones. This means that there is signal processing in the eye before information is sent to the brain. For example, the eye makes comparisons between adjacent light receptors and is very sensitive to edges as seen in [link]. Rather than responding simply to the light entering the eye, which is uniform in the various rectangles in this figure, the eye responds to the edges and senses false darkness variations.



The importance of edges is

shown. Although the grey strips are uniformly shaded, as indicated by the graph immediately below them, they do not appear uniform at all. Instead, they are perceived darker on the dark side and lighter on the light side of the edge, as shown in the bottom graph. This is due to nerve impulse processing in the eye.

One theory that takes various factors into account was advanced by Edwin Land (1909 – 1991), the creative founder of the Polaroid Corporation. Land proposed, based partly on his many elegant experiments, that the three types of cones are organized into systems called **retinexes**. Each retinex forms an image that is compared with the others, and the eye-brain system thus can compare a candle-illuminated white table cloth with its generally reddish surroundings and determine that it is actually white. This **retinex theory of color vision** is an example of modified theories of color vision that attempt to account for its subtleties. One striking experiment performed by Land demonstrates that some type of image comparison may produce color

vision. Two pictures are taken of a scene on black-and-white film, one using a red filter, the other a blue filter. Resulting black-and-white slides are then projected and superimposed on a screen, producing a black-and-white image, as expected. Then a red filter is placed in front of the slide taken with a red filter, and the images are again superimposed on a screen. You would expect an image in various shades of pink, but instead, the image appears to humans in full color with all the hues of the original scene. This implies that color vision can be induced by comparison of the black-and-white and red images. Color vision is not completely understood or explained, and the retinex theory is not totally accepted. It is apparent that color vision is much subtler than what a first look might imply.

Note:
PhET Explorations: Color Vision
Make a whole rainbow by mixing red, green, and blue light. Change the wavelength of a monochromatic beam or filter white light. View the light as a solid beam, or see the individual photons.

https://phet.colorado.edu/sims/html/color-vision/latest/color-vision_en.html

Section Summary

- The eye has four types of light receptors—rods and three types of color-sensitive cones.
- The rods are good for night vision, peripheral vision, and motion changes, while the cones are responsible for central vision and color.
- We perceive many hues, from light having mixtures of wavelengths.
- A simplified theory of color vision states that there are three primary colors, which correspond to the three types of cones, and that various combinations of the primary colors produce all the hues.
- The true color of an object is related to its relative absorption of various wavelengths of light. The color of a light source is related to the wavelengths it produces.

- Color constancy is the ability of the eye-brain system to discern the true color of an object illuminated by various light sources.
- The retinex theory of color vision explains color constancy by postulating the existence of three retinexes or image systems, associated with the three types of cones that are compared to obtain sophisticated information.

Conceptual Questions

Exercise:

Problem:

A pure red object on a black background seems to disappear when illuminated with pure green light. Explain why.

Exercise:

Problem: What is color constancy, and what are its limitations?

Exercise:

Problem:

There are different types of color blindness related to the malfunction of different types of cones. Why would it be particularly useful to study those rare individuals who are color blind only in one eye or who have a different type of color blindness in each eye?

Exercise:

Problem:

Propose a way to study the function of the rods alone, given they can sense light about 1000 times dimmer than the cones.

Glossary

hues

identity of a color as it relates specifically to the spectrum

rods and cones
: two types of photoreceptors in the human retina; rods are responsible for vision at low light levels, while cones are active at higher light levels

simplified theory of color vision
: a theory that states that there are three primary colors, which correspond to the three types of cones

color constancy
: a part of the visual perception system that allows people to perceive color in a variety of conditions and to see some consistency in the color

retinex
: a theory proposed to explain color and brightness perception and constancies; is a combination of the words retina and cortex, which are the two areas responsible for the processing of visual information

retinex theory of color vision
: the ability to perceive color in an ambient-colored environment

Microscopes

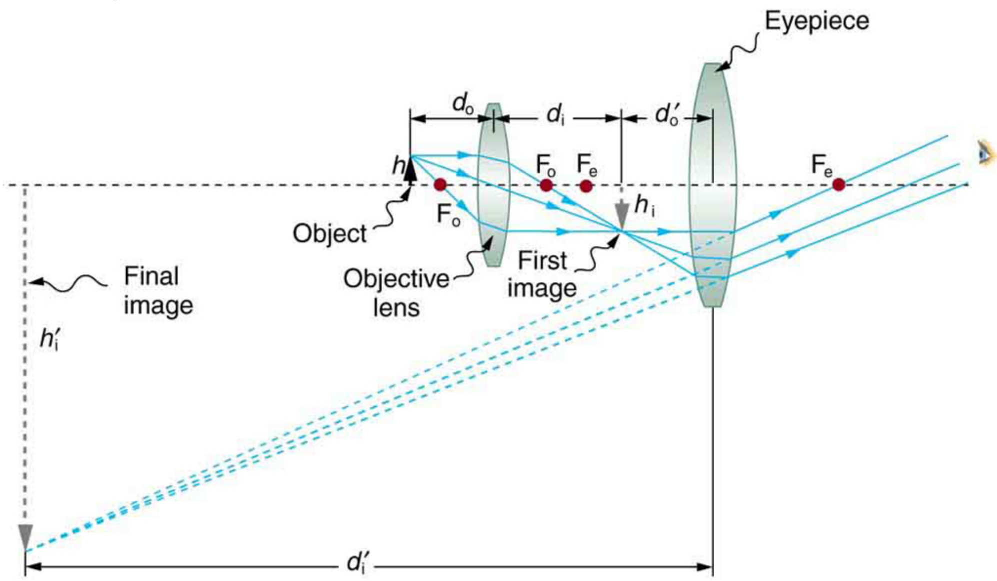
- Investigate different types of microscopes.
- Learn how image is formed in a compound microscope.

Although the eye is marvelous in its ability to see objects large and small, it obviously has limitations to the smallest details it can detect. Human desire to see beyond what is possible with the naked eye led to the use of optical instruments. In this section we will examine microscopes, instruments for enlarging the detail that we cannot see with the unaided eye. The microscope is a multiple-element system having more than a single lens or mirror. (See [link]) A microscope can be made from two convex lenses. The image formed by the first element becomes the object for the second element. The second element forms its own image, which is the object for the third element, and so on. Ray tracing helps to visualize the image formed. If the device is composed of thin lenses and mirrors that obey the thin lens equations, then it is not difficult to describe their behavior numerically.

Multiple lenses and mirrors are used in this microscope. (credit: U.S. Navy photo by Tom Watanabe)

Microscopes were first developed in the early 1600s by eyeglass makers in The Netherlands and Denmark. The simplest **compound microscope** is

constructed from two convex lenses as shown schematically in [link]. The first lens is called the **objective lens**, and has typical magnification values from $5\times$ to $100\times$. In standard microscopes, the objectives are mounted such that when you switch between objectives, the sample remains in focus. Objectives arranged in this way are described as parfocal. The second, the **eyepiece**, also referred to as the ocular, has several lenses which slide inside a cylindrical barrel. The focusing ability is provided by the movement of both the objective lens and the eyepiece. The purpose of a microscope is to magnify small objects, and both lenses contribute to the final magnification. Additionally, the final enlarged image is produced in a location far enough from the observer to be easily viewed, since the eye cannot focus on objects or images that are too close.



A compound microscope composed of two lenses, an objective and an eyepiece. The objective forms a case 1 image that is larger than the object. This first image is the object for the eyepiece. The eyepiece forms a case 2 final image that is further magnified.

To see how the microscope in [link] forms an image, we consider its two lenses in succession. The object is slightly farther away from the objective lens than its focal length f_o, producing a case 1 image that is larger than the

object. This first image is the object for the second lens, or eyepiece. The eyepiece is intentionally located so it can further magnify the image. The eyepiece is placed so that the first image is closer to it than its focal length f_e. Thus the eyepiece acts as a magnifying glass, and the final image is made even larger. The final image remains inverted, but it is farther from the observer, making it easy to view (the eye is most relaxed when viewing distant objects and normally cannot focus closer than 25 cm). Since each lens produces a magnification that multiplies the height of the image, it is apparent that the overall magnification m is the product of the individual magnifications:

Equation:

$$m = m_o m_e,$$

where m_o is the magnification of the objective and m_e is the magnification of the eyepiece. This equation can be generalized for any combination of thin lenses and mirrors that obey the thin lens equations.

Note:

Overall Magnification

The overall magnification of a multiple-element system is the product of the individual magnifications of its elements.

Example:

Microscope Magnification

Calculate the magnification of an object placed 6.20 mm from a compound microscope that has a 6.00 mm focal length objective and a 50.0 mm focal length eyepiece. The objective and eyepiece are separated by 23.0 cm.

Strategy and Concept

This situation is similar to that shown in [link]. To find the overall magnification, we must find the magnification of the objective, then the magnification of the eyepiece. This involves using the thin lens equation.

Solution

The magnification of the objective lens is given as
Equation:

$$m_{\mathrm{o}} = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}},$$

where d_{o} and d_{i} are the object and image distances, respectively, for the objective lens as labeled in [link]. The object distance is given to be $d_{\mathrm{o}} = 6.20\ \mathrm{mm}$, but the image distance d_{i} is not known. Isolating d_{i}, we have
Equation:

$$\frac{1}{d_{\mathrm{i}}} = \frac{1}{f_{\mathrm{o}}} - \frac{1}{d_{\mathrm{o}}},$$

where f_{o} is the focal length of the objective lens. Substituting known values gives
Equation:

$$\frac{1}{d_{\mathrm{i}}} = \frac{1}{6.00\ \mathrm{mm}} - \frac{1}{6.20\ \mathrm{mm}} = \frac{0.00538}{\mathrm{mm}}.$$

We invert this to find d_{i}:
Equation:

$$d_{\mathrm{i}} = 186\ \mathrm{mm}.$$

Substituting this into the expression for m_{o} gives
Equation:

$$m_{\mathrm{o}} = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = -\frac{186\ \mathrm{mm}}{6.20\ \mathrm{mm}} = -30.0.$$

Now we must find the magnification of the eyepiece, which is given by
Equation:

$$m_{\mathrm{e}} = -\frac{d_{\mathrm{i}}\prime}{d_{\mathrm{o}}\prime},$$

where $d_{\text{i}}\prime$ and $d_{\text{o}}\prime$ are the image and object distances for the eyepiece (see [link]). The object distance is the distance of the first image from the eyepiece. Since the first image is 186 mm to the right of the objective and the eyepiece is 230 mm to the right of the objective, the object distance is $d_{\text{o}}\prime = 230\text{ mm} - 186\text{ mm} = 44.0\text{ mm}$. This places the first image closer to the eyepiece than its focal length, so that the eyepiece will form a case 2 image as shown in the figure. We still need to find the location of the final image $d_{\text{i}}\prime$ in order to find the magnification. This is done as before to obtain a value for $1/d_{\text{i}}\prime$:

Equation:

$$\frac{1}{d_{\text{i}}\prime} = \frac{1}{f_{\text{e}}} - \frac{1}{d_{\text{o}}\prime} = \frac{1}{50.0\text{ mm}} - \frac{1}{44.0\text{ mm}} = -\frac{0.00273}{\text{mm}}.$$

Inverting gives

Equation:

$$d_{\text{i}}\prime = -\frac{\text{mm}}{0.00273} = -367\text{ mm}.$$

The eyepiece's magnification is thus

Equation:

$$m_{\text{e}} = -\frac{d_{\text{i}}\prime}{d_{\text{o}}\prime} = -\frac{-367\text{ mm}}{44.0\text{ mm}} = 8.33.$$

So the overall magnification is

Equation:

$$m = m_{\text{o}}m_{\text{e}} = (-30.0)(8.33) = -250.$$

Discussion

Both the objective and the eyepiece contribute to the overall magnification, which is large and negative, consistent with [link], where the image is seen to be large and inverted. In this case, the image is virtual and inverted, which cannot happen for a single element (case 2 and case 3 images for single elements are virtual and upright). The final image is 367 mm (0.367 m) to the left of the eyepiece. Had the eyepiece been placed farther from

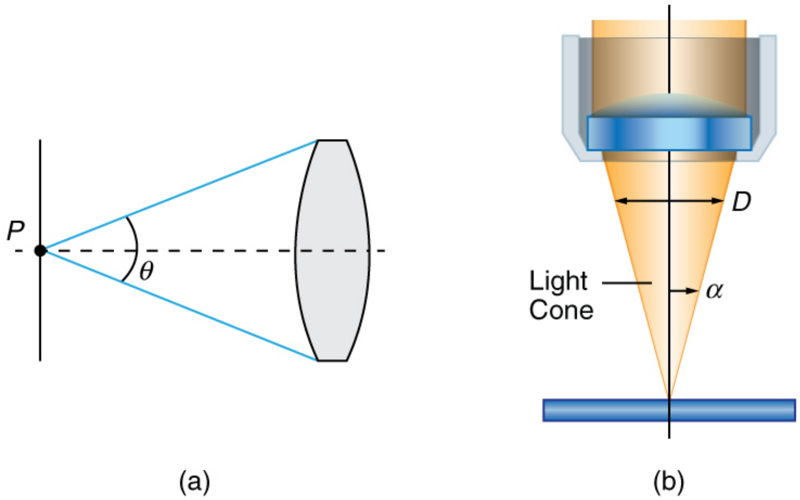
the objective, it could have formed a case 1 image to the right. Such an image could be projected on a screen, but it would be behind the head of the person in the figure and not appropriate for direct viewing. The procedure used to solve this example is applicable in any multiple-element system. Each element is treated in turn, with each forming an image that becomes the object for the next element. The process is not more difficult than for single lenses or mirrors, only lengthier.

Normal optical microscopes can magnify up to $1500\times$ with a theoretical resolution of $-0.2\ \mu\text{m}$. The lenses can be quite complicated and are composed of multiple elements to reduce aberrations. Microscope objective lenses are particularly important as they primarily gather light from the specimen. Three parameters describe microscope objectives: the **numerical aperture** (NA), the magnification (m), and the working distance. The NA is related to the light gathering ability of a lens and is obtained using the angle of acceptance θ formed by the maximum cone of rays focusing on the specimen (see [link](a)) and is given by

Equation:

$$\text{NA} = n\ \text{sin}\ \alpha,$$

where n is the refractive index of the medium between the lens and the specimen and $\alpha = \theta/2$. As the angle of acceptance given by θ increases, NA becomes larger and more light is gathered from a smaller focal region giving higher resolution. A 0.75NA objective gives more detail than a $0.10NA$ objective.



(a) The numerical aperture (NA) of a microscope objective lens refers to the light-gathering ability of the lens and is calculated using half the angle of acceptance θ. (b) Here, α is half the acceptance angle for light rays from a specimen entering a camera lens, and D is the diameter of the aperture that controls the light entering the lens.

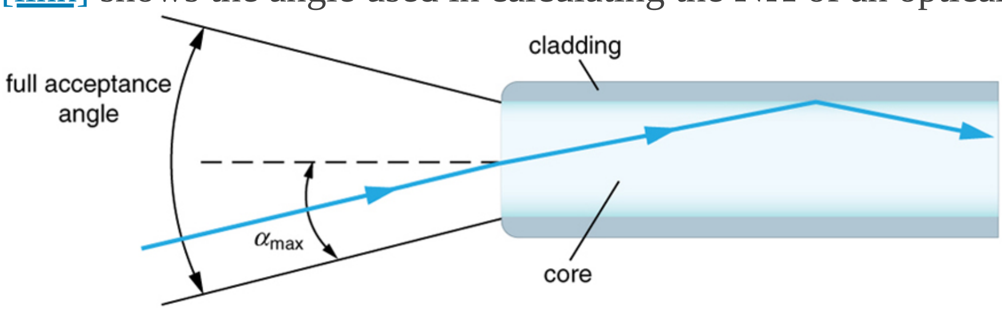
While the numerical aperture can be used to compare resolutions of various objectives, it does not indicate how far the lens could be from the specimen. This is specified by the "working distance," which is the distance (in mm usually) from the front lens element of the objective to the specimen, or cover glass. The higher the NA the closer the lens will be to the specimen and the more chances there are of breaking the cover slip and damaging both the specimen and the lens. The focal length of an objective lens is different than the working distance. This is because objective lenses are made of a combination of lenses and the focal length is measured from inside the barrel. The working distance is a parameter that microscopists can use more readily as it is measured from the outermost lens. The working distance decreases as the NA and magnification both increase.

The term $f/\#$ in general is called the f-number and is used to denote the light per unit area reaching the image plane. In photography, an image of an object at infinity is formed at the focal point and the f-number is given by the ratio of the focal length f of the lens and the diameter D of the aperture controlling the light into the lens (see [link](b)). If the acceptance angle is small the NA of the lens can also be used as given below.

Equation:

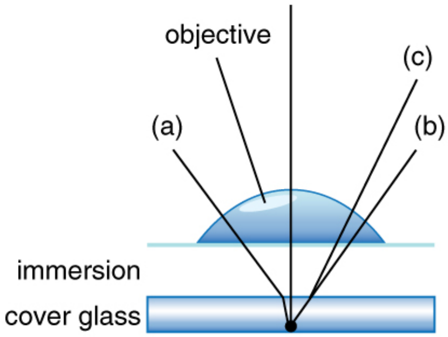
$$f/\# = \frac{f}{D} \approx \frac{1}{2\mathrm{NA}}.$$

As the f-number decreases, the camera is able to gather light from a larger angle, giving wide-angle photography. As usual there is a trade-off. A greater $f/\#$ means less light reaches the image plane. A setting of $f/16$ usually allows one to take pictures in bright sunlight as the aperture diameter is small. In optical fibers, light needs to be focused into the fiber. [link] shows the angle used in calculating the NA of an optical fiber.



Light rays enter an optical fiber. The numerical aperture of the optical fiber can be determined by using the angle α_{max}.

Can the NA be larger than 1.00? The answer is ‘yes’ if we use immersion lenses in which a medium such as oil, glycerine or water is placed between the objective and the microscope cover slip. This minimizes the mismatch in refractive indices as light rays go through different media, generally providing a greater light-gathering ability and an increase in resolution. [link] shows light rays when using air and immersion lenses.

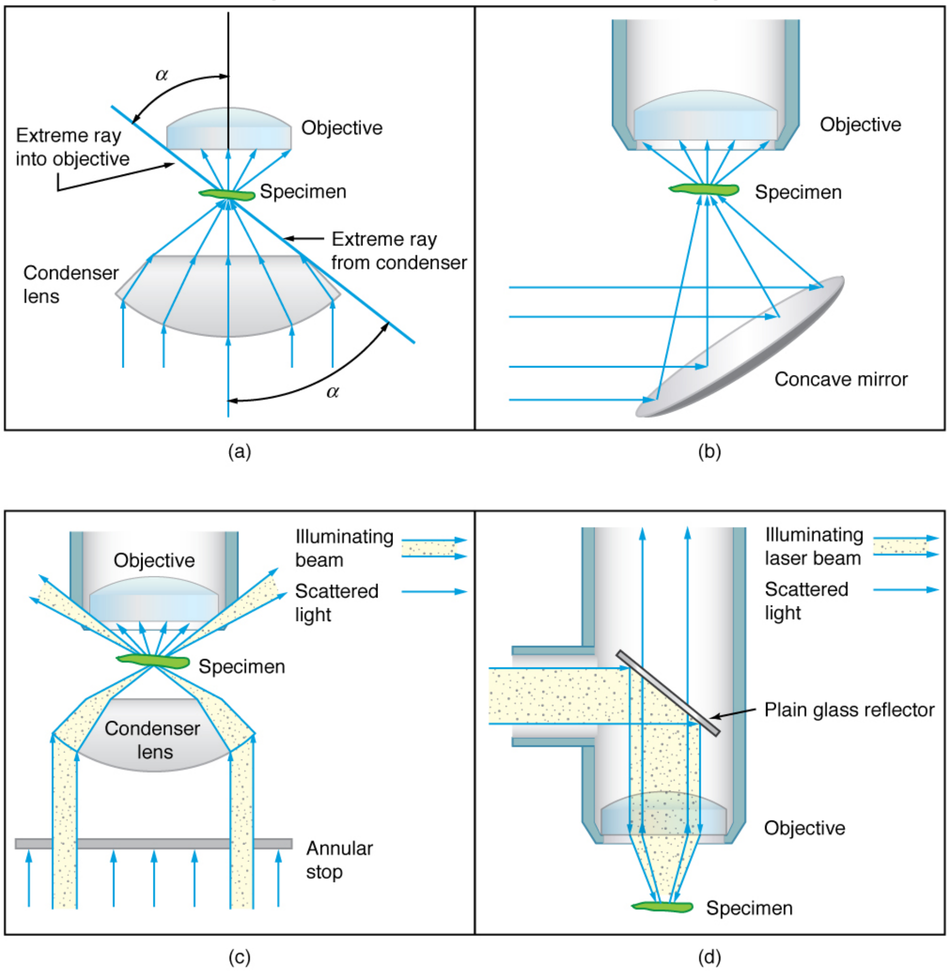


Light rays from a specimen entering the objective. Paths for immersion medium of air (a), water (b) $(n = 1.33)$, and oil (c) $(n = 1.51)$ are shown. The water and oil immersions allow more rays to enter the objective, increasing the resolution.

When using a microscope we do not see the entire extent of the sample. Depending on the eyepiece and objective lens we see a restricted region which we say is the field of view. The objective is then manipulated in two-dimensions above the sample to view other regions of the sample. Electronic scanning of either the objective or the sample is used in scanning microscopy. The image formed at each point during the scanning is combined using a computer to generate an image of a larger region of the sample at a selected magnification.

When using a microscope, we rely on gathering light to form an image. Hence most specimens need to be illuminated, particularly at higher magnifications, when observing details that are so small that they reflect only small amounts of light. To make such objects easily visible, the intensity of light falling on them needs to be increased. Special illuminating

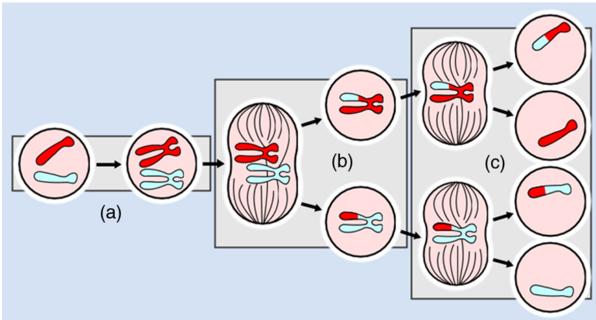
systems called condensers are used for this purpose. The type of condenser that is suitable for an application depends on how the specimen is examined, whether by transmission, scattering or reflecting. See [link] for an example of each. White light sources are common and lasers are often used. Laser light illumination tends to be quite intense and it is important to ensure that the light does not result in the degradation of the specimen.



Illumination of a specimen in a microscope. (a) Transmitted light from a condenser lens. (b) Transmitted light from a mirror condenser. (c) Dark field illumination by scattering (the illuminating beam misses the objective lens). (d) High magnification illumination with reflected light – normally laser light.

We normally associate microscopes with visible light but x ray and electron microscopes provide greater resolution. The focusing and basic physics is the same as that just described, even though the lenses require different technology. The electron microscope requires vacuum chambers so that the electrons can proceed unheeded. Magnifications of 50 million times provide the ability to determine positions of individual atoms within materials. An electron microscope is shown in [link]. We do not use our eyes to form images; rather images are recorded electronically and displayed on computers. In fact observing and saving images formed by optical microscopes on computers is now done routinely. Video recordings of what occurs in a microscope can be made for viewing by many people at later dates. Physics provides the science and tools needed to generate the sequence of time-lapse images of meiosis similar to the sequence sketched in [link].

An electron microscope has the capability to image individual atoms on a material. The microscope uses vacuum technology, sophisticated detectors and state of the art image processing software. (credit: Dave Pape)



The image shows a sequence of events that takes place during meiosis. (credit: PatríciaR, Wikimedia Commons; National Center for Biotechnology Information)

Note:
Take-Home Experiment: Make a Lens
Look through a clear glass or plastic bottle and describe what you see. Now fill the bottle with water and describe what you see. Use the water bottle as a lens to produce the image of a bright object and estimate the focal length of the water bottle lens. How is the focal length a function of the depth of water in the bottle?

Section Summary

- The microscope is a multiple-element system having more than a single lens or mirror.
- Many optical devices contain more than a single lens or mirror. These are analysed by considering each element sequentially. The image formed by the first is the object for the second, and so on. The same ray tracing and thin lens techniques apply to each lens element.

- The overall magnification of a multiple-element system is the product of the magnifications of its individual elements. For a two-element system with an objective and an eyepiece, this is
  Equation:
  $$m = m_{\text{o}} m_{\text{e}},$$
  where m_{o} is the magnification of the objective and m_{e} is the magnification of the eyepiece, such as for a microscope.
- Microscopes are instruments for allowing us to see detail we would not be able to see with the unaided eye and consist of a range of components.
- The eyepiece and objective contribute to the magnification. The numerical aperture (NA) of an objective is given by
  Equation:
  $$\text{NA} = n \sin \alpha$$
  where n is the refractive index and α the angle of acceptance.
- Immersion techniques are often used to improve the light gathering ability of microscopes. The specimen is illuminated by transmitted, scattered or reflected light though a condenser.
- The $f\ /\#$ describes the light gathering ability of a lens. It is given by
  Equation:
  $$f\ /\# = \frac{f}{D} \approx \frac{1}{2\,NA}.$$

Conceptual Questions

Exercise:

Problem:

Geometric optics describes the interaction of light with macroscopic objects. Why, then, is it correct to use geometric optics to analyse a microscope’s image?

Exercise:

Problem:

The image produced by the microscope in [link] cannot be projected. Could extra lenses or mirrors project it? Explain.

Exercise:

Problem:

Why not have the objective of a microscope form a case 2 image with a large magnification? (Hint: Consider the location of that image and the difficulty that would pose for using the eyepiece as a magnifier.)

Exercise:

Problem: What advantages do oil immersion objectives offer?

Exercise:

Problem:

How does the NA of a microscope compare with the NA of an optical fiber?

Problem Exercises

Exercise:

Problem:

A microscope with an overall magnification of 800 has an objective that magnifies by 200. (a) What is the magnification of the eyepiece? (b) If there are two other objectives that can be used, having magnifications of 100 and 400, what other total magnifications are possible?

Solution:

(a) 4.00

(b) 1600

Exercise:

Problem:

(a) What magnification is produced by a 0.150 cm focal length microscope objective that is 0.155 cm from the object being viewed? (b) What is the overall magnification if an $8\times$ eyepiece (one that produces a magnification of 8.00) is used?

Exercise:

Problem:

(a) Where does an object need to be placed relative to a microscope for its 0.500 cm focal length objective to produce a magnification of –400 ? (b) Where should the 5.00 cm focal length eyepiece be placed to produce a further fourfold (4.00) magnification?

Solution:

(a) 0.501 cm

(b) Eyepiece should be 204 cm behind the objective lens.

Exercise:

Problem:

You switch from a $1.40 NA\ 60\times$ oil immersion objective to a $1.40 NA\ 60\times$ oil immersion objective. What are the acceptance angles for each? Compare and comment on the values. Which would you use first to locate the target area on your specimen?

Exercise:

Problem:

An amoeba is 0.305 cm away from the 0.300 cm focal length objective lens of a microscope. (a) Where is the image formed by the objective lens? (b) What is this image's magnification? (c) An eyepiece with a 2.00 cm focal length is placed 20.0 cm from the objective. Where is the final image? (d) What magnification is produced by the eyepiece? (e) What is the overall magnification? (See [link].)

Solution:

(a) +18.3 cm (on the eyepiece side of the objective lens)

(b) -60.0

(c) -11.3 cm (on the objective side of the eyepiece)

(d) +6.67

(e) -400

Exercise:

Problem:

You are using a standard microscope with a $0.10NA$ $4\times$ objective and switch to a $0.65NA$ $40\times$ objective. What are the acceptance angles for each? Compare and comment on the values. Which would you use first to locate the target area on of your specimen? (See [link].)

Exercise:

Problem: Unreasonable Results

Your friends show you an image through a microscope. They tell you that the microscope has an objective with a 0.500 cm focal length and an eyepiece with a 5.00 cm focal length. The resulting overall magnification is 250,000. Are these viable values for a microscope?

Glossary

compound microscope
: a microscope constructed from two convex lenses, the first serving as the ocular lens(close to the eye) and the second serving as the objective lens

objective lens
: the lens nearest to the object being examined

eyepiece
: the lens or combination of lenses in an optical instrument nearest to the eye of the observer

numerical aperture
: a number or measure that expresses the ability of a lens to resolve fine detail in an object being observed. Derived by mathematical formula

Equation:

$$\mathrm{NA} = n \sin \alpha,$$

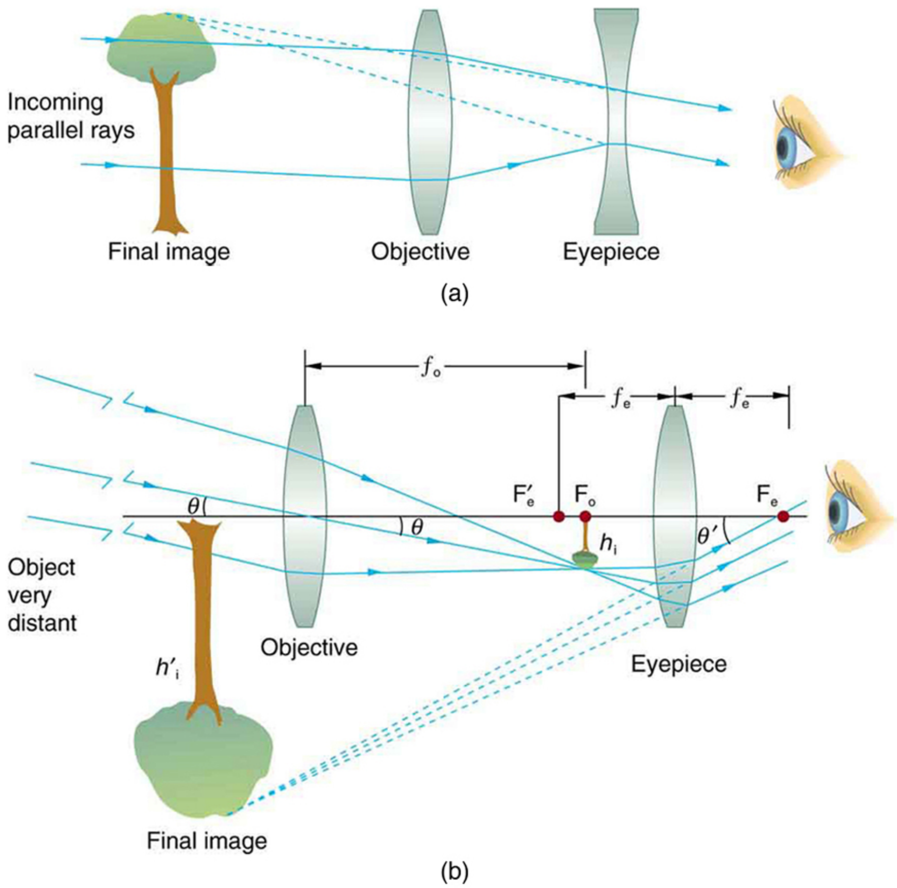
where n is the refractive index of the medium between the lens and the specimen and $\alpha = \theta/2$

Telescopes

- Outline the invention of a telescope.
- Describe the working of a telescope.

Telescopes are meant for viewing distant objects, producing an image that is larger than the image that can be seen with the unaided eye. Telescopes gather far more light than the eye, allowing dim objects to be observed with greater magnification and better resolution. Although Galileo is often credited with inventing the telescope, he actually did not. What he did was more important. He constructed several early telescopes, was the first to study the heavens with them, and made monumental discoveries using them. Among these are the moons of Jupiter, the craters and mountains on the Moon, the details of sunspots, and the fact that the Milky Way is composed of vast numbers of individual stars.

[link](a) shows a telescope made of two lenses, the convex objective and the concave eyepiece, the same construction used by Galileo. Such an arrangement produces an upright image and is used in spyglasses and opera glasses.



(a) Galileo made telescopes with a convex objective and a concave eyepiece. These produce an upright image and are used in spyglasses. (b) Most simple telescopes have two convex lenses. The objective forms a case 1 image that is the object for the eyepiece. The eyepiece forms a case 2 final image that is magnified.

The most common two-lens telescope, like the simple microscope, uses two convex lenses and is shown in [link](b). The object is so far away from the telescope that it is essentially at infinity compared with the focal lengths of the lenses ($d_o \approx \infty$). The first image is thus produced at $d_i = f_o$, as shown in the figure. To prove this, note that

Equation:

$$\frac{1}{d_{\text{i}}} = \frac{1}{f_{\text{o}}} - \frac{1}{d_{\text{o}}} = \frac{1}{f_{\text{o}}} - \frac{1}{\infty}.$$

Because $1/\infty = 0$, this simplifies to
Equation:

$$\frac{1}{d_{\text{i}}} = \frac{1}{f_{\text{o}}},$$

which implies that $d_{\text{i}} = f_{\text{o}}$, as claimed. It is true that for any distant object and any lens or mirror, the image is at the focal length.

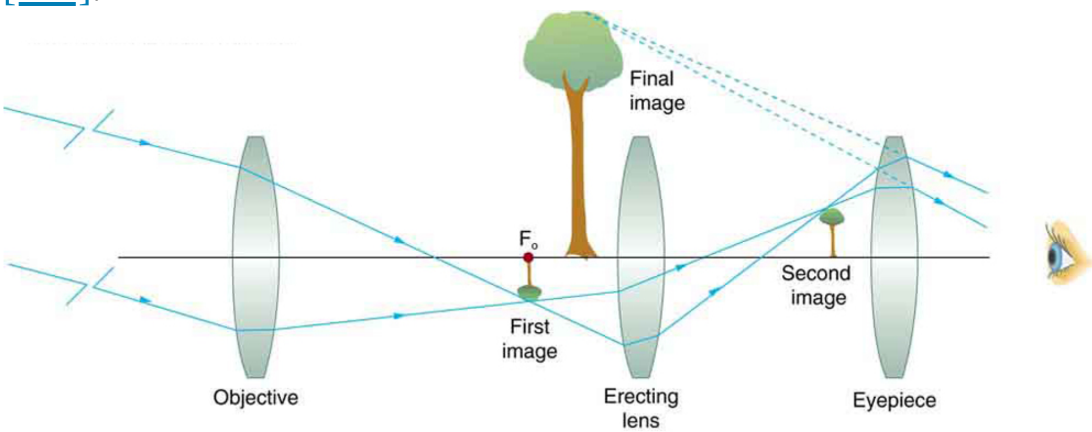
The first image formed by a telescope objective as seen in [link](b) will not be large compared with what you might see by looking at the object directly. For example, the spot formed by sunlight focused on a piece of paper by a magnifying glass is the image of the Sun, and it is small. The telescope eyepiece (like the microscope eyepiece) magnifies this first image. The distance between the eyepiece and the objective lens is made slightly less than the sum of their focal lengths so that the first image is closer to the eyepiece than its focal length. That is, $d_{\text{o}}\prime$ is less than f_{e}, and so the eyepiece forms a case 2 image that is large and to the left for easy viewing. If the angle subtended by an object as viewed by the unaided eye is θ, and the angle subtended by the telescope image is $\theta\prime$, then the **angular magnification** M is defined to be their ratio. That is, $M = \theta\prime/\theta$. It can be shown that the angular magnification of a telescope is related to the focal lengths of the objective and eyepiece; and is given by
Equation:

$$M = \frac{\theta\prime}{\theta} = -\frac{f_{\text{o}}}{f_{\text{e}}}.$$

The minus sign indicates the image is inverted. To obtain the greatest angular magnification, it is best to have a long focal length objective and a short focal length eyepiece. The greater the angular magnification M, the larger an object will appear when viewed through a telescope, making more

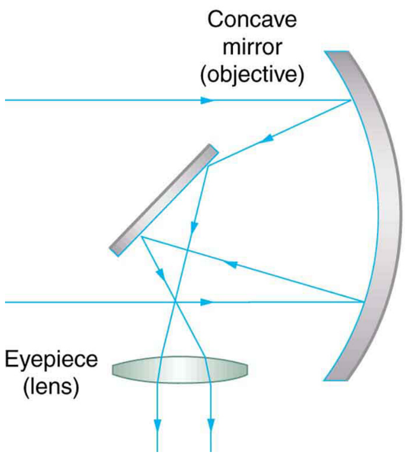
details visible. Limits to observable details are imposed by many factors, including lens quality and atmospheric disturbance.

The image in most telescopes is inverted, which is unimportant for observing the stars but a real problem for other applications, such as telescopes on ships or telescopic gun sights. If an upright image is needed, Galileo's arrangement in [link](a) can be used. But a more common arrangement is to use a third convex lens as an eyepiece, increasing the distance between the first two and inverting the image once again as seen in [link].



This arrangement of three lenses in a telescope produces an upright final image. The first two lenses are far enough apart that the second lens inverts the image of the first one more time. The third lens acts as a magnifier and keeps the image upright and in a location that is easy to view.

A telescope can also be made with a concave mirror as its first element or objective, since a concave mirror acts like a convex lens as seen in [link]. Flat mirrors are often employed in optical instruments to make them more compact or to send light to cameras and other sensing devices. There are many advantages to using mirrors rather than lenses for telescope objectives. Mirrors can be constructed much larger than lenses and can, thus, gather large amounts of light, as needed to view distant galaxies, for example. Large and relatively flat mirrors have very long focal lengths, so that great angular magnification is possible.



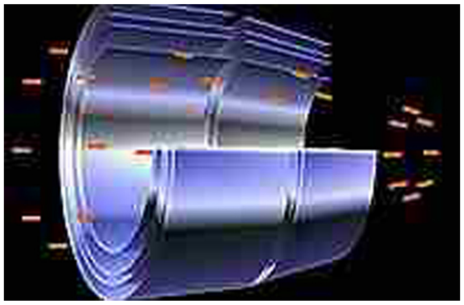
A two-element telescope composed of a mirror as the objective and a lens for the eyepiece is shown. This telescope forms an image in the same manner as the two-convex-lens telescope already discussed, but it does not suffer from chromatic aberrations. Such telescopes can gather more light, since larger mirrors than lenses can be constructed.

Telescopes, like microscopes, can utilize a range of frequencies from the electromagnetic spectrum. [link](a) shows the Australia Telescope Compact

Array, which uses six 22-m antennas for mapping the southern skies using radio waves. [link](b) shows the focusing of x rays on the Chandra X-ray Observatory—a satellite orbiting earth since 1999 and looking at high temperature events as exploding stars, quasars, and black holes. X rays, with much more energy and shorter wavelengths than RF and light, are mainly absorbed and not reflected when incident perpendicular to the medium. But they can be reflected when incident at small glancing angles, much like a rock will skip on a lake if thrown at a small angle. The mirrors for the Chandra consist of a long barrelled pathway and 4 pairs of mirrors to focus the rays at a point 10 meters away from the entrance. The mirrors are extremely smooth and consist of a glass ceramic base with a thin coating of metal (iridium). Four pairs of precision manufactured mirrors are exquisitely shaped and aligned so that x rays ricochet off the mirrors like bullets off a wall, focusing on a spot.

(a)

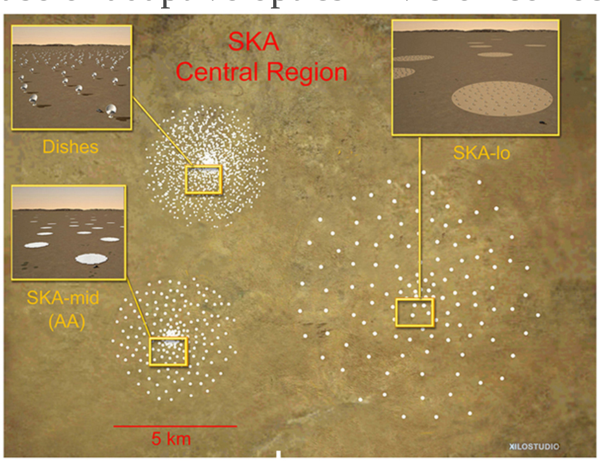
(b)



(a) The Australia Telescope Compact Array at Narrabri (500 km NW of Sydney). (credit: Ian

Bailey) (b) The focusing of x rays on the Chandra Observatory, a satellite orbiting earth. X rays ricochet off 4 pairs of mirrors forming a barrelled pathway leading to the focus point. (credit: NASA)

A current exciting development is a collaborative effort involving 17 countries to construct a Square Kilometre Array (SKA) of telescopes capable of covering from 80 MHz to 2 GHz. The initial stage of the project is the construction of the Australian Square Kilometre Array Pathfinder in Western Australia (see [link]). The project will use cutting-edge technologies such as **adaptive optics** in which the lens or mirror is constructed from lots of carefully aligned tiny lenses and mirrors that can be manipulated using computers. A range of rapidly changing distortions can be minimized by deforming or tilting the tiny lenses and mirrors. The use of adaptive optics in vision correction is a current area of research.



An artist’s impression of the Australian Square Kilometre Array Pathfinder in Western

Australia is displayed. (credit: SPDO, XILOSTUDIOS)

Section Summary

- Simple telescopes can be made with two lenses. They are used for viewing objects at large distances and utilize the entire range of the electromagnetic spectrum.
- The angular magnification M for a telescope is given by
  Equation:

  $$M = \frac{\theta\prime}{\theta} = -\frac{f_{\text{o}}}{f_{\text{e}}},$$

  where θ is the angle subtended by an object viewed by the unaided eye, $\theta\prime$ is the angle subtended by a magnified image, and f_{o} and f_{e} are the focal lengths of the objective and the eyepiece.

Conceptual Questions

Exercise:

Problem:

If you want your microscope or telescope to project a real image onto a screen, how would you change the placement of the eyepiece relative to the objective?

Problem Exercises

Unless otherwise stated, the lens-to-retina distance is 2.00 cm.
Exercise:

Problem:

What is the angular magnification of a telescope that has a 100 cm focal length objective and a 2.50 cm focal length eyepiece?

Solution:

-40.0

Exercise:

Problem:

Find the distance between the objective and eyepiece lenses in the telescope in the above problem needed to produce a final image very far from the observer, where vision is most relaxed. Note that a telescope is normally used to view very distant objects.

Exercise:

Problem:

A large reflecting telescope has an objective mirror with a 10.0 m radius of curvature. What angular magnification does it produce when a 3.00 m focal length eyepiece is used?

Solution:

-1.67

Exercise:

Problem:

A small telescope has a concave mirror with a 2.00 m radius of curvature for its objective. Its eyepiece is a 4.00 cm focal length lens. (a) What is the telescope's angular magnification? (b) What angle is subtended by a 25,000 km diameter sunspot? (c) What is the angle of its telescopic image?

Exercise:

Problem:

A $7.5\times$ binocular produces an angular magnification of -7.50, acting like a telescope. (Mirrors are used to make the image upright.) If the binoculars have objective lenses with a 75.0 cm focal length, what is the focal length of the eyepiece lenses?

Solution:

$+10.0$ cm

Exercise:

Problem: Construct Your Own Problem

Consider a telescope of the type used by Galileo, having a convex objective and a concave eyepiece as illustrated in [link](a). Construct a problem in which you calculate the location and size of the image produced. Among the things to be considered are the focal lengths of the lenses and their relative placements as well as the size and location of the object. Verify that the angular magnification is greater than one. That is, the angle subtended at the eye by the image is greater than the angle subtended by the object.

Glossary

adaptive optics
: optical technology in which computers adjust the lenses and mirrors in a device to correct for image distortions

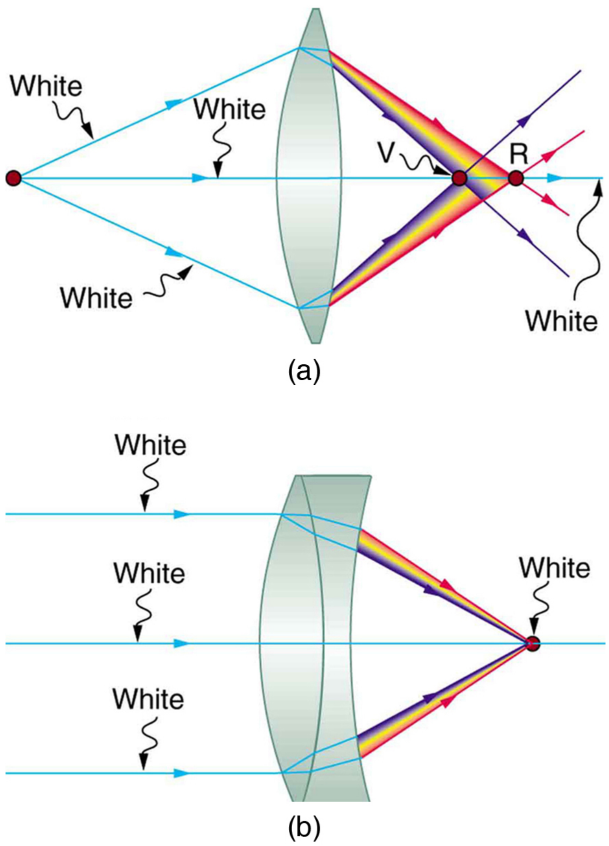
angular magnification
: a ratio related to the focal lengths of the objective and eyepiece and given as $M = -\frac{f_o}{f_e}$

Aberrations

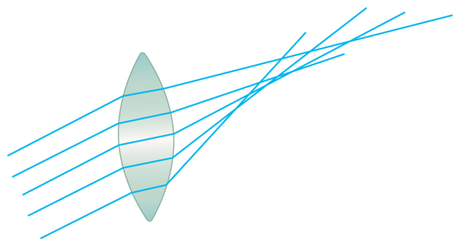
- Describe optical aberration.

Real lenses behave somewhat differently from how they are modeled using the thin lens equations, producing **aberrations**. An aberration is a distortion in an image. There are a variety of aberrations due to a lens size, material, thickness, and position of the object. One common type of aberration is chromatic aberration, which is related to color. Since the index of refraction of lenses depends on color or wavelength, images are produced at different places and with different magnifications for different colors. (The law of reflection is independent of wavelength, and so mirrors do not have this problem. This is another advantage for mirrors in optical systems such as telescopes.) [link](a) shows chromatic aberration for a single convex lens and its partial correction with a two-lens system. Violet rays are bent more than red, since they have a higher index of refraction and are thus focused closer to the lens. The diverging lens partially corrects this, although it is usually not possible to do so completely. Lenses of different materials and having different dispersions may be used. For example an achromatic doublet consisting of a converging lens made of crown glass and a diverging lens made of flint glass in contact can dramatically reduce chromatic aberration (see [link](b)).

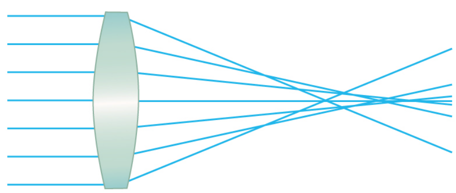
Quite often in an imaging system the object is off-center. Consequently, different parts of a lens or mirror do not refract or reflect the image to the same point. This type of aberration is called a coma and is shown in [link]. The image in this case often appears pear-shaped. Another common aberration is spherical aberration where rays converging from the outer edges of a lens converge to a focus closer to the lens and rays closer to the axis focus further (see [link]). Aberrations due to astigmatism in the lenses of the eyes are discussed in Vision Correction, and a chart used to detect astigmatism is shown in [link]. Such aberrations and can also be an issue with manufactured lenses.



(a) Chromatic aberration is caused by the dependence of a lens's index of refraction on color (wavelength). The lens is more powerful for violet (V) than for red (R), producing images with different locations and magnifications. (b) Multiple-lens systems can partially correct chromatic aberrations, but they may require lenses of different materials and add to the expense of optical systems such as cameras.



A coma is an aberration caused by an object that is off-center, often resulting in a pear-shaped image. The rays originate from points that are not on the optical axis and they do not converge at one common focal point.



Spherical aberration is caused by rays focusing at different distances from the lens.

The image produced by an optical system needs to be bright enough to be discerned. It is often a challenge to obtain a sufficiently bright image. The brightness is determined by the amount of light passing through the optical system. The optical components determining the brightness are the diameter of the lens and the diameter of pupils, diaphragms or aperture stops placed

in front of lenses. Optical systems often have entrance and exit pupils to specifically reduce aberrations but they inevitably reduce brightness as well. Consequently, optical systems need to strike a balance between the various components used. The iris in the eye dilates and constricts, acting as an entrance pupil. You can see objects more clearly by looking through a small hole made with your hand in the shape of a fist. Squinting, or using a small hole in a piece of paper, also will make the object sharper.

So how are aberrations corrected? The lenses may also have specially shaped surfaces, as opposed to the simple spherical shape that is relatively easy to produce. Expensive camera lenses are large in diameter, so that they can gather more light, and need several elements to correct for various aberrations. Further, advances in materials science have resulted in lenses with a range of refractive indices—technically referred to as graded index (GRIN) lenses. Spectacles often have the ability to provide a range of focusing ability using similar techniques. GRIN lenses are particularly important at the end of optical fibers in endoscopes. Advanced computing techniques allow for a range of corrections on images after the image has been collected and certain characteristics of the optical system are known. Some of these techniques are sophisticated versions of what are available on commercial packages like Adobe Photoshop.

Section Summary

- Aberrations or image distortions can arise due to the finite thickness of optical instruments, imperfections in the optical components, and limitations on the ways in which the components are used.
- The means for correcting aberrations range from better components to computational techniques.

Conceptual Questions

Exercise:

Problem:

List the various types of aberrations. What causes them and how can each be reduced?

Problem Exercises

Exercise:

Problem: **Integrated Concepts**

(a) During laser vision correction, a brief burst of 193 nm ultraviolet light is projected onto the cornea of the patient. It makes a spot 1.00 mm in diameter and deposits 0.500 mJ of energy. Calculate the depth of the layer ablated, assuming the corneal tissue has the same properties as water and is initially at 34.0°C. The tissue's temperature is increased to 100°C and evaporated without further temperature increase.

(b) Does your answer imply that the shape of the cornea can be finely controlled?

Solution:

(a) $0.251\ \mu\text{m}$

(b) Yes, this thickness implies that the shape of the cornea can be very finely controlled, producing normal distant vision in more than 90% of patients.

Glossary

aberration
: failure of rays to converge at one focus because of limitations or defects in a lens or mirror

Introduction to Electric Current, Resistance, and Ohm's Law class="introduction"

Electric energy in massive quantities is transmitted from this hydroelectric facility, the Srisailam power station located along the Krishna River in India, by the movement of charge—that is, by electric current. (credit: Chintohere, Wikimedia Commons)

The flicker of numbers on a handheld calculator, nerve impulses carrying signals of vision to the brain, an ultrasound device sending a signal to a computer screen, the brain sending a message for a baby to twitch its toes, an electric train pulling its load over a mountain pass, a hydroelectric plant sending energy to metropolitan and rural users—these and many other examples of electricity involve *electric current, the movement of charge*. Humankind has indeed harnessed electricity, the basis of technology, to improve our quality of life. Whereas the previous two chapters concentrated on static electricity and the fundamental force underlying its behavior, the next few chapters will be devoted to electric and magnetic phenomena involving current. In addition to exploring applications of electricity, we shall gain new insights into nature—in particular, the fact that all magnetism results from electric current.

Current

- Define electric current, ampere, and drift velocity
- Describe the direction of charge flow in conventional current.
- Use drift velocity to calculate current and vice versa.

Electric Current

Electric current is defined to be the rate at which charge flows. A large current, such as that used to start a truck engine, moves a large amount of charge in a small time, whereas a small current, such as that used to operate a hand-held calculator, moves a small amount of charge over a long period of time. In equation form, **electric current** I is defined to be

Equation:

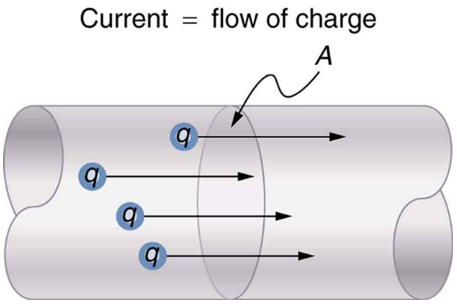
$$I = \frac{\Delta Q}{\Delta t},$$

where ΔQ is the amount of charge passing through a given area in time Δt. (As in previous chapters, initial time is often taken to be zero, in which case $\Delta t = t$.) (See [link].) The SI unit for current is the **ampere** (A), named for the French physicist André-Marie Ampère (1775–1836). Since $I = \Delta Q / \Delta t$, we see that an ampere is one coulomb per second:

Equation:

$$1 \text{ A} = 1 \text{ C/s}$$

Not only are fuses and circuit breakers rated in amperes (or amps), so are many electrical appliances.



The rate of flow of charge is current. An ampere is the flow of one coulomb through an area in one second.

Example:
Calculating Currents: Current in a Truck Battery and a Handheld Calculator

(a) What is the current involved when a truck battery sets in motion 720 C of charge in 4.00 s while starting an engine? (b) How long does it take 1.00 C of charge to flow through a handheld calculator if a 0.300-mA current is flowing?

Strategy

We can use the definition of current in the equation $I = \Delta Q/\Delta t$ to find the current in part (a), since charge and time are given. In part (b), we rearrange the definition of current and use the given values of charge and current to find the time required.

Solution for (a)

Entering the given values for charge and time into the definition of current gives

Equation:

$$\begin{aligned} I &= \frac{\Delta Q}{\Delta t} = \frac{720\text{ C}}{4.00\text{ s}} = 180\text{ C/s} \\ &= 180\text{ A.} \end{aligned}$$

Discussion for (a)

This large value for current illustrates the fact that a large charge is moved in a small amount of time. The currents in these “starter motors” are fairly large because large frictional forces need to be overcome when setting something in motion.

Solution for (b)

Solving the relationship $I = \Delta Q/\Delta t$ for time Δt, and entering the known values for charge and current gives

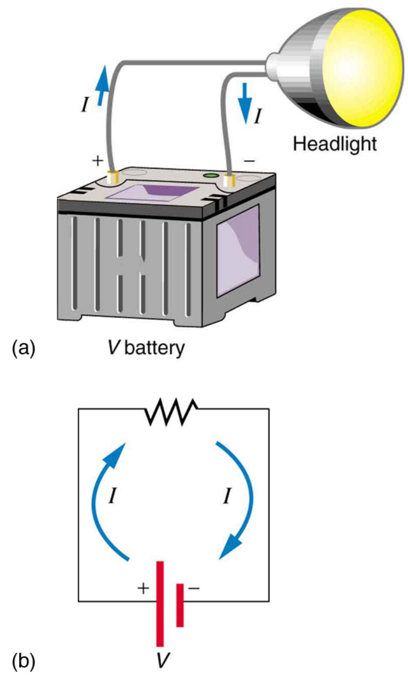
Equation:

$$\begin{aligned}\Delta t &= \frac{\Delta Q}{I} = \frac{1.00\text{ C}}{0.300 \times 10^{-3}\text{ C/s}} \\ &= 3.33 \times 10^{3}\text{ s.}\end{aligned}$$

Discussion for (b)

This time is slightly less than an hour. The small current used by the hand-held calculator takes a much longer time to move a smaller charge than the large current of the truck starter. So why can we operate our calculators only seconds after turning them on? It’s because calculators require very little energy. Such small current and energy demands allow handheld calculators to operate from solar cells or to get many hours of use out of small batteries. Remember, calculators do not have moving parts in the same way that a truck engine has with cylinders and pistons, so the technology requires smaller currents.

[link] shows a simple circuit and the standard schematic representation of a battery, conducting path, and load (a resistor). Schematics are very useful in visualizing the main features of a circuit. A single schematic can represent a wide variety of situations. The schematic in [link] (b), for example, can represent anything from a truck battery connected to a headlight lighting the street in front of the truck to a small battery connected to a penlight lighting a keyhole in a door. Such schematics are useful because the analysis is the same for a wide variety of situations. We need to understand a few schematics to apply the concepts and analysis to many more situations.



(a) A simple electric circuit. A closed path for current to flow through is supplied by conducting wires connecting a load to the terminals of a battery. (b) In this schematic, the battery is represented by the two parallel red lines, conducting wires are shown as straight lines, and the zigzag represents the load. The schematic represents a wide

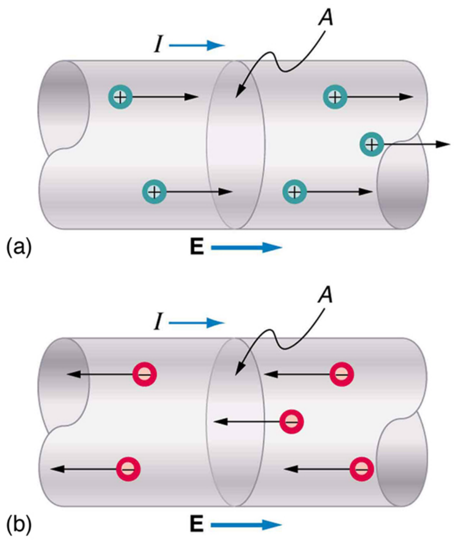
variety of similar circuits.

Note that the direction of current flow in [link] is from positive to negative. *The direction of conventional current is the direction that positive charge would flow*. Depending on the situation, positive charges, negative charges, or both may move. In metal wires, for example, current is carried by electrons—that is, negative charges move. In ionic solutions, such as salt water, both positive and negative charges move. This is also true in nerve cells. A Van de Graaff generator used for nuclear research can produce a current of pure positive charges, such as protons. [link] illustrates the movement of charged particles that compose a current. The fact that conventional current is taken to be in the direction that positive charge would flow can be traced back to American politician and scientist Benjamin Franklin in the 1700s. He named the type of charge associated with electrons negative, long before they were known to carry current in so many situations. Franklin, in fact, was totally unaware of the small-scale structure of electricity.

It is important to realize that there is an electric field in conductors responsible for producing the current, as illustrated in [link]. Unlike static electricity, where a conductor in equilibrium cannot have an electric field in it, conductors carrying a current have an electric field and are not in static equilibrium. An electric field is needed to supply energy to move the charges.

Note:
Making Connections: Take-Home Investigation—Electric Current Illustration
Find a straw and little peas that can move freely in the straw. Place the straw flat on a table and fill the straw with peas. When you pop one pea in at one end, a different pea should pop out the other end. This demonstration is an analogy for an electric current. Identify what compares

to the electrons and what compares to the supply of energy. What other analogies can you find for an electric current?
Note that the flow of peas is based on the peas physically bumping into each other; electrons flow due to mutually repulsive electrostatic forces.



Current I is the rate at which charge moves through an area A, such as the cross-section of a wire. Conventional current is defined to move in the direction of the electric field. (a) Positive charges move in the direction of the electric field and the same direction as conventional current. (b) Negative charges move in the direction opposite to the electric field. Conventional

current is in the direction opposite to the movement of negative charge. The flow of electrons is sometimes referred to as electronic flow.

Example:

Calculating the Number of Electrons that Move through a Calculator

If the 0.300-mA current through the calculator mentioned in the [link] example is carried by electrons, how many electrons per second pass through it?

Strategy

The current calculated in the previous example was defined for the flow of positive charge. For electrons, the magnitude is the same, but the sign is opposite, $I_{\text{electrons}} = -0.300 \times 10^{-3}\ \text{C/s}$ .Since each electron (e^{-}) has a charge of $–1.60 \times 10^{-19}$ C, we can convert the current in coulombs per second to electrons per second.

Solution

Starting with the definition of current, we have

Equation:

$$I_{\text{electrons}} = \frac{\Delta Q_{\text{electrons}}}{\Delta t} = \frac{–0.300 \times 10^{-3}\ \text{C}}{\text{s}}.$$

We divide this by the charge per electron, so that

Equation:

$$\begin{aligned}\frac{e^{-}}{\text{s}} &= \frac{–0.300\times 10^{-3}\ \text{C}}{\text{s}} \times \frac{1\ e^{-}}{–1.60\times 10^{-19}\ \text{C}}\\ &= 1.88\times 10^{15}\ \frac{e^{-}}{\text{s}}.\end{aligned}$$

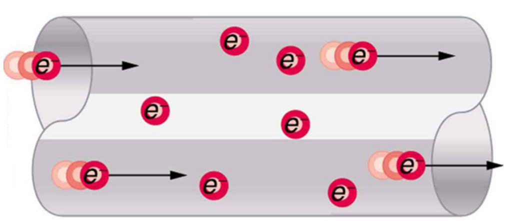
Discussion

There are so many charged particles moving, even in small currents, that individual charges are not noticed, just as individual water molecules are not noticed in water flow. Even more amazing is that they do not always keep moving forward like soldiers in a parade. Rather they are like a crowd of people with movement in different directions but a general trend to move forward. There are lots of collisions with atoms in the metal wire and, of course, with other electrons.

Drift Velocity

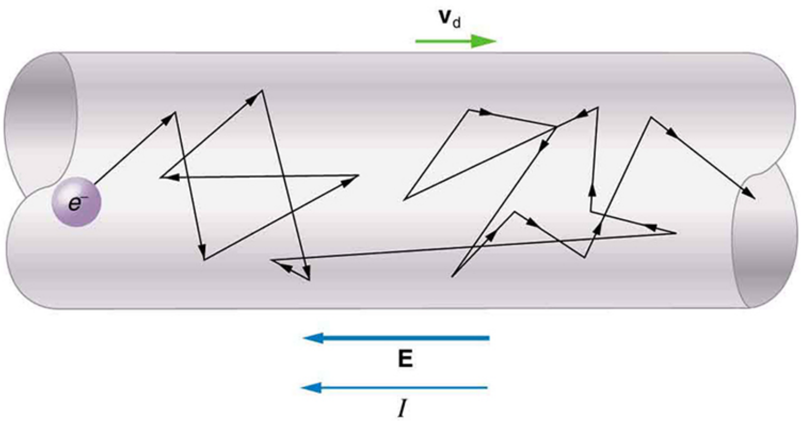
Electrical signals are known to move very rapidly. Telephone conversations carried by currents in wires cover large distances without noticeable delays. Lights come on as soon as a switch is flicked. Most electrical signals carried by currents travel at speeds on the order of $10^8\ \text{m/s}$, a significant fraction of the speed of light. Interestingly, the individual charges that make up the current move *much* more slowly on average, typically drifting at speeds on the order of $10^{-4}\ \text{m/s}$. How do we reconcile these two speeds, and what does it tell us about standard conductors?

The high speed of electrical signals results from the fact that the force between charges acts rapidly at a distance. Thus, when a free charge is forced into a wire, as in [link], the incoming charge pushes other charges ahead of it, which in turn push on charges farther down the line. The density of charge in a system cannot easily be increased, and so the signal is passed on rapidly. The resulting electrical shock wave moves through the system at nearly the speed of light. To be precise, this rapidly moving signal or shock wave is a rapidly propagating change in electric field.



When charged particles are forced into this volume of a conductor, an equal number are quickly forced to leave. The repulsion between like charges makes it difficult to increase the number of charges in a volume. Thus, as one charge enters, another leaves almost immediately, carrying the signal rapidly forward.

Good conductors have large numbers of free charges in them. In metals, the free charges are free electrons. [link] shows how free electrons move through an ordinary conductor. The distance that an individual electron can move between collisions with atoms or other electrons is quite small. The electron paths thus appear nearly random, like the motion of atoms in a gas. But there is an electric field in the conductor that causes the electrons to drift in the direction shown (opposite to the field, since they are negative). The **drift velocity** v_{d} is the average velocity of the free charges. Drift velocity is quite small, since there are so many free charges. If we have an estimate of the density of free electrons in a conductor, we can calculate the drift velocity for a given current. The larger the density, the lower the velocity required for a given current.



Free electrons moving in a conductor make many collisions with other electrons and atoms. The path of one electron is shown. The average velocity of the free charges is called the drift velocity, v_{d}, and it is in the direction opposite to the electric field for electrons. The collisions normally transfer energy to the conductor, requiring a constant supply of energy to maintain a steady current.

Note:
Conduction of Electricity and Heat
Good electrical conductors are often good heat conductors, too. This is because large numbers of free electrons can carry electrical current and can transport thermal energy.

The free-electron collisions transfer energy to the atoms of the conductor. The electric field does work in moving the electrons through a distance, but that work does not increase the kinetic energy (nor speed, therefore) of the electrons. The work is transferred to the conductor’s atoms, possibly

increasing temperature. Thus a continuous power input is required to keep a current flowing. An exception, of course, is found in superconductors, for reasons we shall explore in a later chapter. Superconductors can have a steady current without a continual supply of energy—a great energy savings. In contrast, the supply of energy can be useful, such as in a lightbulb filament. The supply of energy is necessary to increase the temperature of the tungsten filament, so that the filament glows.

Note:
Making Connections: Take-Home Investigation—Filament Observations
Find a lightbulb with a filament. Look carefully at the filament and describe its structure. To what points is the filament connected?

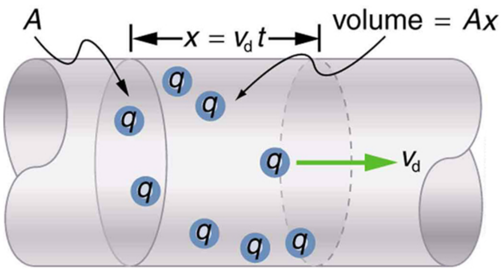
We can obtain an expression for the relationship between current and drift velocity by considering the number of free charges in a segment of wire, as illustrated in [link]. *The number of free charges per unit volume* is given the symbol n and depends on the material. The shaded segment has a volume Ax, so that the number of free charges in it is nAx. The charge ΔQ in this segment is thus qnAx, where q is the amount of charge on each carrier. (Recall that for electrons, q is $-1.60 \times 10^{-19}\ \mathrm{C}$.) Current is charge moved per unit time; thus, if all the original charges move out of this segment in time Δt, the current is
Equation:

$$I = \frac{\Delta Q}{\Delta t} = \frac{\mathrm{qnAx}}{\Delta t}.$$

Note that $x/\Delta t$ is the magnitude of the drift velocity, v_{d}, since the charges move an average distance x in a time Δt. Rearranging terms gives
Equation:

$$I = \mathrm{nqAv}_{\mathrm{d}},$$

where I is the current through a wire of cross-sectional area A made of a material with a free charge density n. The carriers of the current each have charge q and move with a drift velocity of magnitude v_{d}.



All the charges in the shaded volume of this wire move out in a time t, having a drift velocity of magnitude $v_{\mathrm{d}} = x/t$. See text for further discussion.

Note that simple drift velocity is not the entire story. The speed of an electron is much greater than its drift velocity. In addition, not all of the electrons in a conductor can move freely, and those that do might move somewhat faster or slower than the drift velocity. So what do we mean by free electrons? Atoms in a metallic conductor are packed in the form of a lattice structure. Some electrons are far enough away from the atomic nuclei that they do not experience the attraction of the nuclei as much as the inner electrons do. These are the free electrons. They are not bound to a single atom but can instead move freely among the atoms in a “sea” of electrons. These free electrons respond by accelerating when an electric field is applied. Of course as they move they collide with the atoms in the lattice and other electrons, generating thermal energy, and the conductor gets warmer. In an insulator, the organization of the atoms and the structure do not allow for such free electrons.

Example:
Calculating Drift Velocity in a Common Wire
Calculate the drift velocity of electrons in a 12-gauge copper wire (which has a diameter of 2.053 mm) carrying a 20.0-A current, given that there is one free electron per copper atom. (Household wiring often contains 12-gauge copper wire, and the maximum current allowed in such wire is usually 20 A.) The density of copper is $8.80 \times 10^3\ \text{kg/m}^3$.
Strategy
We can calculate the drift velocity using the equation $I = nqAv_\text{d}$. The current $I = 20.0\ \text{A}$ is given, and $q = -1.60 \times 10^{-19}\text{C}$ is the charge of an electron. We can calculate the area of a cross-section of the wire using the formula $A = \pi r^2$, where r is one-half the given diameter, 2.053 mm. We are given the density of copper, $8.80 \times 10^3\ \text{kg/m}^3$, and the periodic table shows that the atomic mass of copper is 63.54 g/mol. We can use these two quantities along with Avogadro's number, $6.02 \times 10^{23}\ \text{atoms/mol}$, to determine n, the number of free electrons per cubic meter.
Solution
First, calculate the density of free electrons in copper. There is one free electron per copper atom. Therefore, is the same as the number of copper atoms per m^3. We can now find n as follows:
Equation:

$$\begin{aligned} n &= \frac{1\ e^-}{\text{atom}} \times \frac{6.02\times10^{23}\ \text{atoms}}{\text{mol}} \times \frac{1\ \text{mol}}{63.54\ \text{g}} \times \frac{1000\ \text{g}}{\text{kg}} \times \frac{8.80\times10^{3}\ \text{kg}}{1\ \text{m}^3} \\ &= 8.342 \times 10^{28}\ e^-/\text{m}^3. \end{aligned}$$

The cross-sectional area of the wire is
Equation:

$$\begin{aligned} A &= \pi r^2 \\ &= \pi\left(\frac{2.053\times10^{-3}\ \text{m}}{2}\right)^2 \\ &= 3.310 \times 10^{-6}\ \text{m}^2. \end{aligned}$$

Rearranging $I = nqAv_\text{d}$ to isolate drift velocity gives
Equation:

$$v_d = \frac{I}{nqA}$$
$$= \frac{20.0\text{ A}}{(8.342\times10^{28}/\text{m}^3)(-1.60\times10^{-19}\text{ C})(3.310\times10^{-6}\text{ m}^2)}$$
$$= -4.53 \times 10^{-4}\text{ m/s}.$$

Discussion

The minus sign indicates that the negative charges are moving in the direction opposite to conventional current. The small value for drift velocity (on the order of 10^{-4} m/s) confirms that the signal moves on the order of 10^{12} times faster (about 10^8 m/s) than the charges that carry it.

Section Summary

- Electric current I is the rate at which charge flows, given by **Equation:**

  $$I = \frac{\Delta Q}{\Delta t},$$

  where ΔQ is the amount of charge passing through an area in time Δt.
- The direction of conventional current is taken as the direction in which positive charge moves.
- The SI unit for current is the ampere (A), where $1\text{ A} = 1\text{ C/s}$.
- Current is the flow of free charges, such as electrons and ions.
- Drift velocity v_d is the average speed at which these charges move.
- Current I is proportional to drift velocity v_d, as expressed in the relationship $I = \text{nqAv}_d$. Here, I is the current through a wire of cross-sectional area A. The wire's material has a free-charge density n, and each carrier has charge q and a drift velocity v_d.
- Electrical signals travel at speeds about 10^{12} times greater than the drift velocity of free electrons.

Conceptual Questions

Exercise:

Problem:

Can a wire carry a current and still be neutral—that is, have a total charge of zero? Explain.

Exercise:

Problem:

Car batteries are rated in ampere-hours ($\mathrm{A \cdot h}$). To what physical quantity do ampere-hours correspond (voltage, charge, . . .), and what relationship do ampere-hours have to energy content?

Exercise:

Problem:

If two different wires having identical cross-sectional areas carry the same current, will the drift velocity be higher or lower in the better conductor? Explain in terms of the equation $v_{\mathrm{d}} = \frac{I}{\mathrm{nqA}}$, by considering how the density of charge carriers n relates to whether or not a material is a good conductor.

Exercise:

Problem:

Why are two conducting paths from a voltage source to an electrical device needed to operate the device?

Exercise:

Problem:

In cars, one battery terminal is connected to the metal body. How does this allow a single wire to supply current to electrical devices rather than two wires?

Exercise:

Problem:

Why isn't a bird sitting on a high-voltage power line electrocuted? Contrast this with the situation in which a large bird hits two wires simultaneously with its wings.

Problems & Exercises

Exercise:

Problem:

What is the current in milliamperes produced by the solar cells of a pocket calculator through which 4.00 C of charge passes in 4.00 h?

Solution:

0.278 mA

Exercise:

Problem:

A total of 600 C of charge passes through a flashlight in 0.500 h. What is the average current?

Exercise:

Problem:

What is the current when a typical static charge of $0.250\ \mu\text{C}$ moves from your finger to a metal doorknob in $1.00\ \mu\text{s}$?

Solution:

0.250 A

Exercise:

Problem:

Find the current when 2.00 nC jumps between your comb and hair over a 0.500 - μs time interval.

Exercise:

Problem:

A large lightning bolt had a 20,000-A current and moved 30.0 C of charge. What was its duration?

Solution:

1.50ms

Exercise:

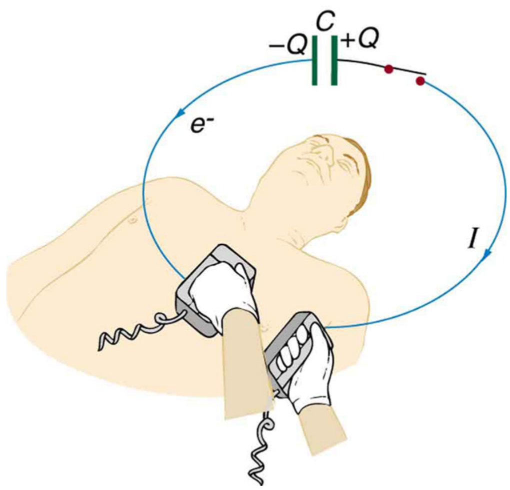
Problem:

The 200-A current through a spark plug moves 0.300 mC of charge. How long does the spark last?

Exercise:

Problem:

(a) A defibrillator sends a 6.00-A current through the chest of a patient by applying a 10,000-V potential as in the figure below. What is the resistance of the path? (b) The defibrillator paddles make contact with the patient through a conducting gel that greatly reduces the path resistance. Discuss the difficulties that would ensue if a larger voltage were used to produce the same current through the patient, but with the path having perhaps 50 times the resistance. (Hint: The current must be about the same, so a higher voltage would imply greater power. Use this equation for power: $P = I^2R$.)



The capacitor in a defibrillation unit drives a current through the heart of a patient.

Solution:

(a) $1.67 \text{k}\Omega$

(b) If a 50 times larger resistance existed, keeping the current about the same, the power would be increased by a factor of about 50 (based on the equation $P = I^2R$), causing much more energy to be transferred to the skin, which could cause serious burns. The gel used reduces the resistance, and therefore reduces the power transferred to the skin.

Exercise:

Problem:

During open-heart surgery, a defibrillator can be used to bring a patient out of cardiac arrest. The resistance of the path is $500\ \Omega$ and a 10.0-mA current is needed. What voltage should be applied?

Exercise:

Problem:

(a) A defibrillator passes 12.0 A of current through the torso of a person for 0.0100 s. How much charge moves? (b) How many electrons pass through the wires connected to the patient? (See figure two problems earlier.)

Solution:

(a) 0.120 C

(b) $7.50 \times 10^{17} \text{electrons}$

Exercise:

Problem:

A clock battery wears out after moving 10,000 C of charge through the clock at a rate of 0.500 mA. (a) How long did the clock run? (b) How many electrons per second flowed?

Exercise:

Problem:

The batteries of a submerged non-nuclear submarine supply 1000 A at full speed ahead. How long does it take to move Avogadro's number (6.02×10^{23}) of electrons at this rate?

Solution:

96.3 s

Exercise:

Problem:

Electron guns are used in X-ray tubes. The electrons are accelerated through a relatively large voltage and directed onto a metal target, producing X-rays. (a) How many electrons per second strike the target if the current is 0.500 mA? (b) What charge strikes the target in 0.750 s?

Exercise:

Problem:

A large cyclotron directs a beam of $\mathrm{He^{++}}$ nuclei onto a target with a beam current of 0.250 mA. (a) How many $\mathrm{He^{++}}$ nuclei per second is this? (b) How long does it take for 1.00 C to strike the target? (c) How long before 1.00 mol of $\mathrm{He^{++}}$ nuclei strike the target?

Solution:

(a) $7.81 \times 10^{14}\ \mathrm{He^{++}\ nuclei/s}$

(b) $4.00 \times 10^{3}\ \mathrm{s}$

(c) $7.71 \times 10^{8}\ \mathrm{s}$

Exercise:

Problem:

Repeat the above example on [link], but for a wire made of silver and given there is one free electron per silver atom.

Exercise:

Problem:

Using the results of the above example on [link], find the drift velocity in a copper wire of twice the diameter and carrying 20.0 A.

Solution:

$-1.13 \times 10^{-4} \mathrm{m/s}$

Exercise:

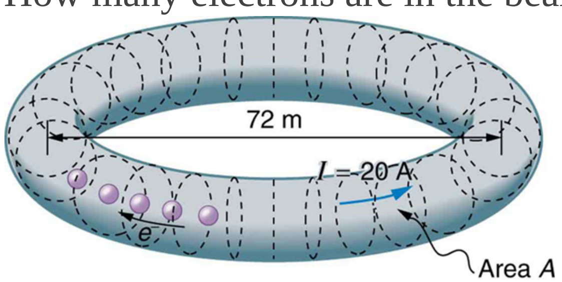
Problem:

A 14-gauge copper wire has a diameter of 1.628 mm. What magnitude current flows when the drift velocity is 1.00 mm/s? (See above example on [link] for useful information.)

Exercise:

Problem:

SPEAR, a storage ring about 72.0 m in diameter at the Stanford Linear Accelerator (closed in 2009), has a 20.0-A circulating beam of electrons that are moving at nearly the speed of light. (See [link].) How many electrons are in the beam?



Electrons circulating in the storage ring called SPEAR constitute a 20.0-A current. Because they travel close to the speed of light, each electron completes many orbits in each second.

Solution:

$9.42 \times 10^{13} \mathrm{electrons}$

Glossary

electric current
: the rate at which charge flows, $I = \Delta Q / \Delta t$

ampere
: (amp) the SI unit for current; 1 A = 1 C/s

drift velocity
: the average velocity at which free charges flow in response to an electric field

Ohm’s Law: Resistance and Simple Circuits

- Explain the origin of Ohm’s law.
- Calculate voltages, currents, or resistances with Ohm’s law.
- Explain what an ohmic material is.
- Describe a simple circuit.

What drives current? We can think of various devices—such as batteries, generators, wall outlets, and so on—which are necessary to maintain a current. All such devices create a potential difference and are loosely referred to as voltage sources. When a voltage source is connected to a conductor, it applies a potential difference V that creates an electric field. The electric field in turn exerts force on charges, causing current.

Ohm’s Law

The current that flows through most substances is directly proportional to the voltage V applied to it. The German physicist Georg Simon Ohm (1787–1854) was the first to demonstrate experimentally that the current in a metal wire is *directly proportional to the voltage applied*:
Equation:

$$I \propto V.$$

This important relationship is known as **Ohm’s law**. It can be viewed as a cause-and-effect relationship, with voltage the cause and current the effect. This is an empirical law like that for friction—an experimentally observed phenomenon. Such a linear relationship doesn’t always occur.

Resistance and Simple Circuits

If voltage drives current, what impedes it? The electric property that impedes current (crudely similar to friction and air resistance) is called **resistance** R. Collisions of moving charges with atoms and molecules in a substance transfer energy to the substance and limit current. Resistance is defined as inversely proportional to current, or

Equation:

$$I \propto \frac{1}{R}.$$

Thus, for example, current is cut in half if resistance doubles. Combining the relationships of current to voltage and current to resistance gives

Equation:

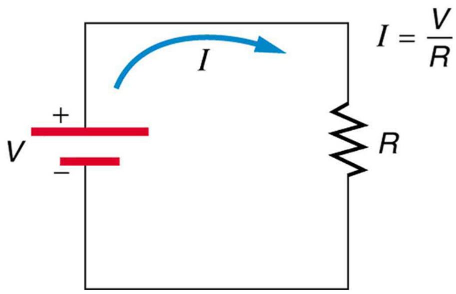
$$I = \frac{V}{R}.$$

This relationship is also called Ohm's law. Ohm's law in this form really defines resistance for certain materials. Ohm's law (like Hooke's law) is not universally valid. The many substances for which Ohm's law holds are called **ohmic**. These include good conductors like copper and aluminum, and some poor conductors under certain circumstances. Ohmic materials have a resistance R that is independent of voltage V and current I. An object that has simple resistance is called a *resistor*, even if its resistance is small. The unit for resistance is an **ohm** and is given the symbol Ω (upper case Greek omega). Rearranging $I = \mathrm{V/R}$ gives $R = \mathrm{V/I}$, and so the units of resistance are 1 ohm = 1 volt per ampere:

Equation:

$$1\ \Omega = 1\frac{V}{A}.$$

[link] shows the schematic for a simple circuit. A **simple circuit** has a single voltage source and a single resistor. The wires connecting the voltage source to the resistor can be assumed to have negligible resistance, or their resistance can be included in R.



A simple electric circuit in which a closed path for current to flow is supplied by conductors (usually metal wires) connecting a load to the terminals of a battery, represented by the red parallel lines. The zigzag symbol represents the single resistor and includes any resistance in the connections to the voltage source.

Example:
Calculating Resistance: An Automobile Headlight
What is the resistance of an automobile headlight through which 2.50 A flows when 12.0 V is applied to it?
Strategy
We can rearrange Ohm's law as stated by $I = \mathrm{V}/\mathrm{R}$ and use it to find the resistance.
Solution
Rearranging $I = \mathrm{V}/\mathrm{R}$ and substituting known values gives

Equation:

$$R = \frac{V}{I} = \frac{12.0\ \text{V}}{2.50\ \text{A}} = 4.80\ \Omega.$$

Discussion

This is a relatively small resistance, but it is larger than the cold resistance of the headlight. As we shall see in Resistance and Resistivity, resistance usually increases with temperature, and so the bulb has a lower resistance when it is first switched on and will draw considerably more current during its brief warm-up period.

Resistances range over many orders of magnitude. Some ceramic insulators, such as those used to support power lines, have resistances of $10^{12}\ \Omega$ or more. A dry person may have a hand-to-foot resistance of $10^{5}\ \Omega$, whereas the resistance of the human heart is about $10^{3}\ \Omega$. A meter-long piece of large-diameter copper wire may have a resistance of $10^{-5}\ \Omega$, and superconductors have no resistance at all (they are non-ohmic). Resistance is related to the shape of an object and the material of which it is composed, as will be seen in Resistance and Resistivity.

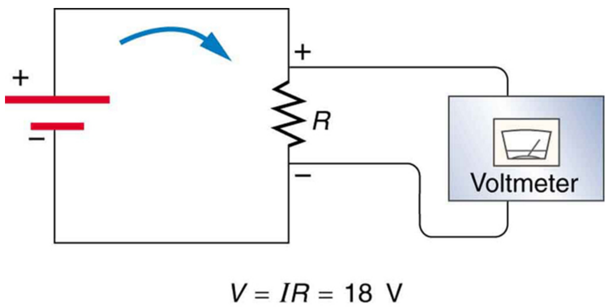
Additional insight is gained by solving $I = \text{V/R}$ for V, yielding

Equation:

$$V = \text{IR}.$$

This expression for V can be interpreted as the *voltage drop across a resistor produced by the flow of current* I. The phrase IR *drop* is often used for this voltage. For instance, the headlight in [link] has an IR drop of 12.0 V. If voltage is measured at various points in a circuit, it will be seen to increase at the voltage source and decrease at the resistor. Voltage is similar to fluid pressure. The voltage source is like a pump, creating a pressure difference, causing current—the flow of charge. The resistor is like a pipe that reduces pressure and limits flow because of its resistance. Conservation of energy has important consequences here. The voltage source supplies

energy (causing an electric field and a current), and the resistor converts it to another form (such as thermal energy). In a simple circuit (one with a single simple resistor), the voltage supplied by the source equals the voltage drop across the resistor, since $\text{PE} = q\Delta V$, and the same q flows through each. Thus the energy supplied by the voltage source and the energy converted by the resistor are equal. (See [link].)



The voltage drop across a resistor in a simple circuit equals the voltage output of the battery.

Note:
Making Connections: Conservation of Energy
In a simple electrical circuit, the sole resistor converts energy supplied by the source into another form. Conservation of energy is evidenced here by the fact that all of the energy supplied by the source is converted to another form by the resistor alone. We will find that conservation of energy has other important applications in circuits and is a powerful tool in circuit analysis.

Note:

PhET Explorations: Ohm's Law
See how the equation form of Ohm's law relates to a simple circuit. Adjust the voltage and resistance, and see the current change according to Ohm's law. The sizes of the symbols in the equation change to match the circuit diagram.
https://phet.colorado.edu/sims/html/ohms-law/latest/ohms-law_en.html

Section Summary

- A simple circuit *is* one in which there is a single voltage source and a single resistance.
- One statement of Ohm's law gives the relationship between current I, voltage V, and resistance R in a simple circuit to be $I=\frac{V}{R}$.
- Resistance has units of ohms (Ω), related to volts and amperes by $1\ \Omega=1\ \text{V/A}$.
- There is a voltage or IR drop across a resistor, caused by the current flowing through it, given by $V=\text{IR}$.

Conceptual Questions

Exercise:

Problem:

The IR drop across a resistor means that there is a change in potential or voltage across the resistor. Is there any change in current as it passes through a resistor? Explain.

Exercise:

Problem:

How is the IR drop in a resistor similar to the pressure drop in a fluid flowing through a pipe?

Problems & Exercises

Exercise:

Problem:

What current flows through the bulb of a 3.00-V flashlight when its hot resistance is $3.60\ \Omega$?

Solution:

0.833 A

Exercise:

Problem:

Calculate the effective resistance of a pocket calculator that has a 1.35-V battery and through which 0.200 mA flows.

Exercise:

Problem:

What is the effective resistance of a car's starter motor when 150 A flows through it as the car battery applies 11.0 V to the motor?

Solution:

$7.33 \times 10^{-2}\ \Omega$

Exercise:

Problem:

How many volts are supplied to operate an indicator light on a DVD player that has a resistance of $140\ \Omega$, given that 25.0 mA passes through it?

Exercise:

Problem:

(a) Find the voltage drop in an extension cord having a $0.0600\text{-}\Omega$ resistance and through which 5.00 A is flowing. (b) A cheaper cord utilizes thinner wire and has a resistance of $0.300\ \Omega$. What is the voltage drop in it when 5.00 A flows? (c) Why is the voltage to whatever appliance is being used reduced by this amount? What is the effect on the appliance?

Solution:

(a) 0.300 V

(b) 1.50 V

(c) The voltage supplied to whatever appliance is being used is reduced because the total voltage drop from the wall to the final output of the appliance is fixed. Thus, if the voltage drop across the extension cord is large, the voltage drop across the appliance is significantly decreased, so the power output by the appliance can be significantly decreased, reducing the ability of the appliance to work properly.

Exercise:

Problem:

A power transmission line is hung from metal towers with glass insulators having a resistance of $1.00\times10^{9}\ \Omega$. What current flows through the insulator if the voltage is 200 kV? (Some high-voltage lines are DC.)

Glossary

Ohm's law
: an empirical relation stating that the current *I* is proportional to the potential difference *V*, $\propto V$; it is often written as $I = V/R$, where *R* is the resistance

resistance

the electric property that impedes current; for ohmic materials, it is the ratio of voltage to current, $R = V/I$

ohm

the unit of resistance, given by $1\Omega = 1\ \text{V/A}$

ohmic

a type of a material for which Ohm's law is valid

simple circuit

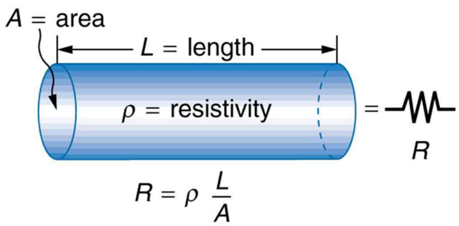
a circuit with a single voltage source and a single resistor

Resistance and Resistivity

- Explain the concept of resistivity.
- Use resistivity to calculate the resistance of specified configurations of material.
- Use the thermal coefficient of resistivity to calculate the change of resistance with temperature.

Material and Shape Dependence of Resistance

The resistance of an object depends on its shape and the material of which it is composed. The cylindrical resistor in [link] is easy to analyze, and, by so doing, we can gain insight into the resistance of more complicated shapes. As you might expect, the cylinder's electric resistance R is directly proportional to its length L, similar to the resistance of a pipe to fluid flow. The longer the cylinder, the more collisions charges will make with its atoms. The greater the diameter of the cylinder, the more current it can carry (again similar to the flow of fluid through a pipe). In fact, R is inversely proportional to the cylinder's cross-sectional area A.



A uniform cylinder of length L and cross-sectional area A. Its resistance to the flow of current is similar to the resistance posed by a pipe to fluid flow. The longer the cylinder, the greater its

resistance. The larger its cross-sectional area A, the smaller its resistance.

For a given shape, the resistance depends on the material of which the object is composed. Different materials offer different resistance to the flow of charge. We define the **resistivity** ρ of a substance so that the **resistance** R of an object is directly proportional to ρ. Resistivity ρ is an *intrinsic* property of a material, independent of its shape or size. The resistance R of a uniform cylinder of length L, of cross-sectional area A, and made of a material with resistivity ρ, is

Equation:

$$R = \frac{\rho L}{A}.$$

[link] gives representative values of ρ. The materials listed in the table are separated into categories of conductors, semiconductors, and insulators, based on broad groupings of resistivities. Conductors have the smallest resistivities, and insulators have the largest; semiconductors have intermediate resistivities. Conductors have varying but large free charge densities, whereas most charges in insulators are bound to atoms and are not free to move. Semiconductors are intermediate, having far fewer free charges than conductors, but having properties that make the number of free charges depend strongly on the type and amount of impurities in the semiconductor. These unique properties of semiconductors are put to use in modern electronics, as will be explored in later chapters.

Material	**Resistivity ρ ($\Omega \cdot \mathrm{m}$)**
Conductors	
Silver	1.59×10^{-8}
Copper	1.72×10^{-8}
Gold	2.44×10^{-8}
Aluminum	2.65×10^{-8}
Tungsten	5.6×10^{-8}
Iron	9.71×10^{-8}
Platinum	10.6×10^{-8}
Steel	20×10^{-8}
Lead	22×10^{-8}

Material	**Resistivity ρ ($\Omega \cdot \mathrm{m}$)**
Manganin (Cu, Mn, Ni alloy)	44×10^{-8}
Constantan (Cu, Ni alloy)	49×10^{-8}
Mercury	96×10^{-8}
Nichrome (Ni, Fe, Cr alloy)	100×10^{-8}
Semiconductors[footnote] Values depend strongly on amounts and types of impurities	
Carbon (pure)	3.5×10^{-5}
Carbon	$(3.5 - 60) \times 10^{-5}$
Germanium (pure)	600×10^{-3}
Germanium	$(1 - 600) \times 10^{-3}$

Material	**Resistivity ρ ($\Omega \cdot \mathrm{m}$)**
Silicon (pure)	2300
Silicon	$0.1–2300$
Insulators	
Amber	5×10^{14}
Glass	$10^9 - 10^{14}$
Lucite	$>10^{13}$
Mica	$10^{11} - 10^{15}$
Quartz (fused)	75×10^{16}
Rubber (hard)	$10^{13} - 10^{16}$
Sulfur	10^{15}

Material	Resistivity ρ ($\Omega \cdot \mathrm{m}$)
Teflon	$>10^{13}$
Wood	$10^{8} - 10^{11}$

Resistivities ρ of Various materials at $20^{\circ}\mathrm{C}$

Example:
Calculating Resistor Diameter: A Headlight Filament
A car headlight filament is made of tungsten and has a cold resistance of $0.350\ \Omega$. If the filament is a cylinder 4.00 cm long (it may be coiled to save space), what is its diameter?
Strategy
We can rearrange the equation $R = \frac{\rho L}{A}$ to find the cross-sectional area A of the filament from the given information. Then its diameter can be found by assuming it has a circular cross-section.
Solution
The cross-sectional area, found by rearranging the expression for the resistance of a cylinder given in $R = \frac{\rho L}{A}$, is
Equation:

$$A = \frac{\rho L}{R}.$$

Substituting the given values, and taking ρ from [link], yields
Equation:

$$\begin{aligned} A &= \frac{(5.6\times10^{-8}\ \Omega\cdot\mathrm{m})(4.00\times10^{-2}\ \mathrm{m})}{0.350\ \Omega} \\ &= 6.40\times10^{-9}\ \mathrm{m}^2. \end{aligned}$$

The area of a circle is related to its diameter D by
Equation:

$$A = \frac{\pi D^2}{4}.$$

Solving for the diameter D, and substituting the value found for A, gives
Equation:

$$\begin{aligned} D &= 2\left(\frac{A}{p}\right)^{\frac{1}{2}} = 2\left(\frac{6.40\times10^{-9}\ \text{m}^2}{3.14}\right)^{\frac{1}{2}} \\ &= 9.0\times10^{-5}\ \text{m}. \end{aligned}$$

Discussion
The diameter is just under a tenth of a millimeter. It is quoted to only two digits, because ρ is known to only two digits.

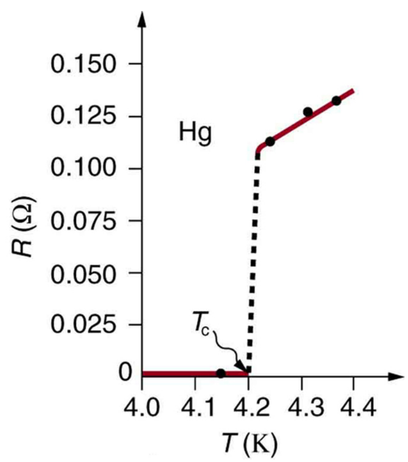
Temperature Variation of Resistance

The resistivity of all materials depends on temperature. Some even become superconductors (zero resistivity) at very low temperatures. (See [link].) Conversely, the resistivity of conductors increases with increasing temperature. Since the atoms vibrate more rapidly and over larger distances at higher temperatures, the electrons moving through a metal make more collisions, effectively making the resistivity higher. Over relatively small temperature changes (about 100ºC or less), resistivity ρ varies with temperature change ΔT as expressed in the following equation
Equation:

$$\rho = \rho_0(1 + \alpha \Delta T),$$

where ρ_0 is the original resistivity and α is the **temperature coefficient of resistivity**. (See the values of α in [link] below.) For larger temperature changes, α may vary or a nonlinear equation may be needed to find ρ. Note

that α is positive for metals, meaning their resistivity increases with temperature. Some alloys have been developed specifically to have a small temperature dependence. Manganin (which is made of copper, manganese and nickel), for example, has α close to zero (to three digits on the scale in [link]), and so its resistivity varies only slightly with temperature. This is useful for making a temperature-independent resistance standard, for example.



The resistance of a sample of mercury is zero at very low temperatures—it is a superconductor up to about 4.2 K. Above that critical temperature, its resistance makes a sudden jump and then increases nearly linearly with temperature.

Material	**Coefficient α(1/°C)[footnote] Values at 20°C.**
Conductors	
Silver	3.8×10^{-3}
Copper	3.9×10^{-3}
Gold	3.4×10^{-3}
Aluminum	3.9×10^{-3}
Tungsten	4.5×10^{-3}
Iron	5.0×10^{-3}
Platinum	3.93×10^{-3}
Lead	3.9×10^{-3}
Manganin (Cu, Mn, Ni alloy)	0.000×10^{-3}

Material	**Coefficient α(1/°C)[footnote] Values at 20°C.**
Constantan (Cu, Ni alloy)	0.002×10^{-3}
Mercury	0.89×10^{-3}
Nichrome (Ni, Fe, Cr alloy)	0.4×10^{-3}
Semiconductors	
Carbon (pure)	-0.5×10^{-3}
Germanium (pure)	-50×10^{-3}
Silicon (pure)	-70×10^{-3}

Tempature Coefficients of Resistivity α

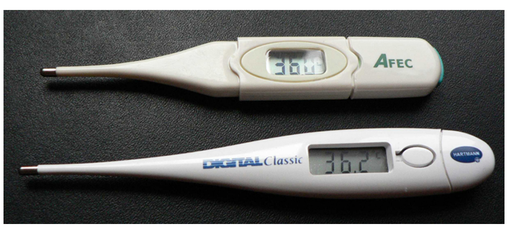
Note also that α is negative for the semiconductors listed in [link], meaning that their resistivity decreases with increasing temperature. They become better conductors at higher temperature, because increased thermal agitation increases the number of free charges available to carry current. This property of decreasing ρ with temperature is also related to the type and amount of impurities present in the semiconductors.

The resistance of an object also depends on temperature, since R_0 is directly proportional to ρ. For a cylinder we know $R = \rho L/A$, and so, if L and A do not change greatly with temperature, R will have the same temperature dependence as ρ. (Examination of the coefficients of linear expansion shows them to be about two orders of magnitude less than typical temperature coefficients of resistivity, and so the effect of temperature on L and A is about two orders of magnitude less than on ρ.) Thus,

Equation:

$$R = R_0(1 + \alpha \Delta T)$$

is the temperature dependence of the resistance of an object, where R_0 is the original resistance and R is the resistance after a temperature change ΔT. Numerous thermometers are based on the effect of temperature on resistance. (See [link].) One of the most common is the thermistor, a semiconductor crystal with a strong temperature dependence, the resistance of which is measured to obtain its temperature. The device is small, so that it quickly comes into thermal equilibrium with the part of a person it touches.



These familiar thermometers are based on the automated measurement of a thermistor’s temperature-dependent resistance. (credit: Biol, Wikimedia Commons)

Example:

Calculating Resistance: Hot-Filament Resistance

Although caution must be used in applying $\rho = \rho_0(1 + \alpha\Delta T)$ and $R = R_0(1 + \alpha\Delta T)$ for temperature changes greater than 100ºC, for tungsten the equations work reasonably well for very large temperature changes. What, then, is the resistance of the tungsten filament in the previous example if its temperature is increased from room temperature (20ºC) to a typical operating temperature of 2850ºC?

Strategy

This is a straightforward application of $R = R_0(1 + \alpha\Delta T)$, since the original resistance of the filament was given to be $R_0 = 0.350\ \Omega$, and the temperature change is $\Delta T = 2830ºC$.

Solution

The hot resistance R is obtained by entering known values into the above equation:

Equation:

$$\begin{aligned} R &= R_0(1 + \alpha\Delta T) \\ &= (0.350\ \Omega)[1 + (4.5\times10^{-3}/ºC)(2830ºC)] \\ &= 4.8\ \Omega. \end{aligned}$$

Discussion

This value is consistent with the headlight resistance example in Ohm’s Law: Resistance and Simple Circuits.

Note:

PhET Explorations: Resistance in a Wire

Learn about the physics of resistance in a wire. Change its resistivity, length, and area to see how they affect the wire's resistance. The sizes of the symbols in the equation change along with the diagram of a wire.

https://phet.colorado.edu/sims/html/resistance-in-a-wire/latest/resistance-in-a-wire_en.html

Section Summary

- The resistance R of a cylinder of length L and cross-sectional area A is $R=\frac{\rho L}{A}$, where ρ is the resistivity of the material.
- Values of ρ in [link] show that materials fall into three groups—*conductors, semiconductors, and insulators*.
- Temperature affects resistivity; for relatively small temperature changes ΔT, resistivity is $\rho=\rho_0(1+\alpha\Delta T)$, where ρ_0 is the original resistivity and α is the temperature coefficient of resistivity.
- [link] gives values for α, the temperature coefficient of resistivity.
- The resistance R of an object also varies with temperature: $R=R_0(1+\alpha\Delta T)$, where R_0 is the original resistance, and R is the resistance after the temperature change.

Conceptual Questions

Exercise:

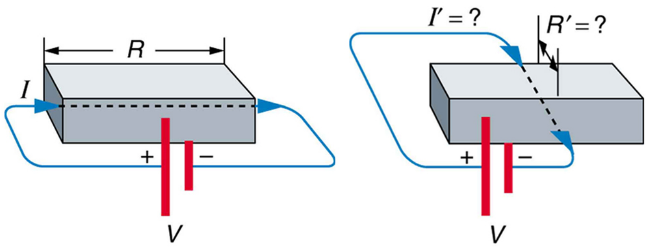
Problem:

In which of the three semiconducting materials listed in [link] do impurities supply free charges? (Hint: Examine the range of resistivity for each and determine whether the pure semiconductor has the higher or lower conductivity.)

Exercise:

Problem:

Does the resistance of an object depend on the path current takes through it? Consider, for example, a rectangular bar—is its resistance the same along its length as across its width? (See [link].)



Does current taking two different paths through the same object encounter different resistance?

Exercise:

Problem:

If aluminum and copper wires of the same length have the same resistance, which has the larger diameter? Why?

Exercise:

Problem:

Explain why $R = R_0(1 + \alpha\Delta T)$ for the temperature variation of the resistance R of an object is not as accurate as $\rho = \rho_0(1 + \alpha\Delta T)$, which gives the temperature variation of resistivity ρ.

Problems & Exercises

Exercise:

Problem:

What is the resistance of a 20.0-m-long piece of 12-gauge copper wire having a 2.053-mm diameter?

Solution:

0.104 Ω

Exercise:

Problem:

The diameter of 0-gauge copper wire is 8.252 mm. Find the resistance of a 1.00-km length of such wire used for power transmission.

Exercise:

Problem:

If the 0.100-mm diameter tungsten filament in a light bulb is to have a resistance of $0.200\ \Omega$ at 20.0°C, how long should it be?

Solution:

$2.8 \times 10^{-2}\ \text{m}$

Exercise:

Problem:

Find the ratio of the diameter of aluminum to copper wire, if they have the same resistance per unit length (as they might in household wiring).

Exercise:

Problem:

What current flows through a 2.54-cm-diameter rod of pure silicon that is 20.0 cm long, when 1.00×10^{3} V is applied to it? (Such a rod may be used to make nuclear-particle detectors, for example.)

Solution:

$1.10 \times 10^{-3}\ \text{A}$

Exercise:

Problem:

(a) To what temperature must you raise a copper wire, originally at $20.0^{\circ}C$, to double its resistance, neglecting any changes in dimensions? (b) Does this happen in household wiring under ordinary circumstances?

Exercise:

Problem:

A resistor made of Nichrome wire is used in an application where its resistance cannot change more than 1.00% from its value at $20.0^{\circ}C$. Over what temperature range can it be used?

Solution:

$-5^{\circ}C$ to $45^{\circ}C$

Exercise:

Problem:

Of what material is a resistor made if its resistance is 40.0% greater at $100^{\circ}C$ than at $20.0^{\circ}C$?

Exercise:

Problem:

An electronic device designed to operate at any temperature in the range from $-10.0^{\circ}C$ to $55.0^{\circ}C$ contains pure carbon resistors. By what factor does their resistance increase over this range?

Solution:

1.03

Exercise:

Problem:

(a) Of what material is a wire made, if it is 25.0 m long with a 0.100 mm diameter and has a resistance of $77.7\ \Omega$ at $20.0^{\circ}C$? (b) What is its resistance at $150^{\circ}C$?

Exercise:

Problem:

Assuming a constant temperature coefficient of resistivity, what is the maximum percent decrease in the resistance of a constantan wire starting at $20.0^{\circ}C$?

Solution:

0.06%

Exercise:

Problem:

A wire is drawn through a die, stretching it to four times its original length. By what factor does its resistance increase?

Exercise:

Problem:

A copper wire has a resistance of $0.500\ \Omega$ at $20.0^{\circ}C$, and an iron wire has a resistance of $0.525\ \Omega$ at the same temperature. At what temperature are their resistances equal?

Solution:

$-17^{\circ}C$

Exercise:

Problem:

(a) Digital medical thermometers determine temperature by measuring the resistance of a semiconductor device called a thermistor (which has $\alpha = -0.0600/^{\circ}\text{C}$) when it is at the same temperature as the patient. What is a patient's temperature if the thermistor's resistance at that temperature is 82.0% of its value at 37.0°C (normal body temperature)? (b) The negative value for α may not be maintained for very low temperatures. Discuss why and whether this is the case here. (Hint: Resistance can't become negative.)

Exercise:

Problem: Integrated Concepts

(a) Redo [link] taking into account the thermal expansion of the tungsten filament. You may assume a thermal expansion coefficient of $12 \times 10^{-6}/^{\circ}\text{C}$. (b) By what percentage does your answer differ from that in the example?

Solution:

(a) $4.7\ \Omega$ (total)

(b) 3.0% decrease

Exercise:

Problem: Unreasonable Results

(a) To what temperature must you raise a resistor made of constantan to double its resistance, assuming a constant temperature coefficient of resistivity? (b) To cut it in half? (c) What is unreasonable about these results? (d) Which assumptions are unreasonable, or which premises are inconsistent?

Glossary

resistivity

an intrinsic property of a material, independent of its shape or size, directly proportional to the resistance, denoted by ρ

temperature coefficient of resistivity

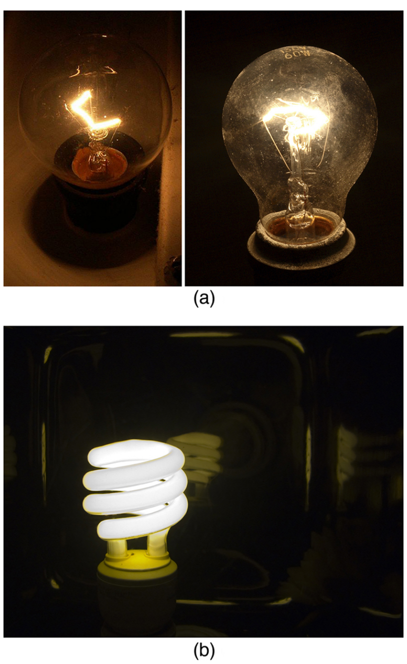
an empirical quantity, denoted by α, which describes the change in resistance or resistivity of a material with temperature

Electric Power and Energy

- Calculate the power dissipated by a resistor and power supplied by a power supply.
- Calculate the cost of electricity under various circumstances.

Power in Electric Circuits

Power is associated by many people with electricity. Knowing that power is the rate of energy use or energy conversion, what is the expression for **electric power**? Power transmission lines might come to mind. We also think of lightbulbs in terms of their power ratings in watts. Let us compare a 25-W bulb with a 60-W bulb. (See [link](a).) Since both operate on the same voltage, the 60-W bulb must draw more current to have a greater power rating. Thus the 60-W bulb's resistance must be lower than that of a 25-W bulb. If we increase voltage, we also increase power. For example, when a 25-W bulb that is designed to operate on 120 V is connected to 240 V, it briefly glows very brightly and then burns out. Precisely how are voltage, current, and resistance related to electric power?



(a) Which of these lightbulbs, the 25-W bulb (upper left) or the 60-W bulb (upper right), has the higher resistance? Which draws more current? Which uses the most energy? Can you tell from the color that the 25-W filament is cooler? Is the brighter bulb a different color and if so why? (credits: Dickbauch, Wikimedia Commons; Greg Westfall, Flickr) (b) This compact fluorescent light (CFL) puts out the same intensity of light as the 60-W bulb, but at 1/4 to 1/10 the input power. (credit: dbgg1979, Flickr)

Electric energy depends on both the voltage involved and the charge moved. This is expressed most simply as $\mathrm{PE} = \mathrm{qV}$, where q is the charge moved and V is the voltage (or more precisely, the potential difference the

charge moves through). Power is the rate at which energy is moved, and so electric power is

Equation:

$$P = \frac{\text{PE}}{t} = \frac{\text{qV}}{t}.$$

Recognizing that current is $I = q/t$ (note that $\Delta t = t$ here), the expression for power becomes

Equation:

$$P = \text{IV}.$$

Electric power (P) is simply the product of current times voltage. Power has familiar units of watts. Since the SI unit for potential energy (PE) is the joule, power has units of joules per second, or watts. Thus, $1\ \text{A} \cdot \text{V} = 1\ \text{W}$. For example, cars often have one or more auxiliary power outlets with which you can charge a cell phone or other electronic devices. These outlets may be rated at 20 A, so that the circuit can deliver a maximum power $P = \text{IV} = (20\ \text{A})(12\ \text{V}) = 240\ \text{W}$. In some applications, electric power may be expressed as volt-amperes or even kilovolt-amperes ($1\ \text{kA} \cdot \text{V} = 1\ \text{kW}$).

To see the relationship of power to resistance, we combine Ohm's law with $P = \text{IV}$. Substituting $I = \text{V/R}$ gives $P = (V/R)V = V^2/R$. Similarly, substituting $V = \text{IR}$ gives $P = I(\text{IR}) = I^2R$. Three expressions for electric power are listed together here for convenience:

Equation:

$$P = \text{IV}$$

Equation:

$$P = \frac{V^2}{R}$$

Equation:

$$P = I^2 R.$$

Note that the first equation is always valid, whereas the other two can be used only for resistors. In a simple circuit, with one voltage source and a single resistor, the power supplied by the voltage source and that dissipated by the resistor are identical. (In more complicated circuits, P can be the power dissipated by a single device and not the total power in the circuit.)

Different insights can be gained from the three different expressions for electric power. For example, $P = V^2/R$ implies that the lower the resistance connected to a given voltage source, the greater the power delivered. Furthermore, since voltage is squared in $P = V^2/R$, the effect of applying a higher voltage is perhaps greater than expected. Thus, when the voltage is doubled to a 25-W bulb, its power nearly quadruples to about 100 W, burning it out. If the bulb's resistance remained constant, its power would be exactly 100 W, but at the higher temperature its resistance is higher, too.

Example:
Calculating Power Dissipation and Current: Hot and Cold Power
(a) Consider the examples given in Ohm's Law: Resistance and Simple Circuits and Resistance and Resistivity. Then find the power dissipated by the car headlight in these examples, both when it is hot and when it is cold.
(b) What current does it draw when cold?
Strategy for (a)
For the hot headlight, we know voltage and current, so we can use $P = \text{IV}$ to find the power. For the cold headlight, we know the voltage and resistance, so we can use $P = V^2/R$ to find the power.
Solution for (a)
Entering the known values of current and voltage for the hot headlight, we obtain
Equation:

$$P = \text{IV} = (2.50\text{ A})(12.0\text{ V}) = 30.0\text{ W}.$$

The cold resistance was $0.350\ \Omega$, and so the power it uses when first switched on is
Equation:

$$P = \frac{V^2}{R} = \frac{(12.0\ \text{V})^2}{0.350\ \Omega} = 411\ \text{W}.$$

Discussion for (a)
The 30 W dissipated by the hot headlight is typical. But the 411 W when cold is surprisingly higher. The initial power quickly decreases as the bulb's temperature increases and its resistance increases.
Strategy and Solution for (b)
The current when the bulb is cold can be found several different ways. We rearrange one of the power equations, $P = I^2R$, and enter known values, obtaining
Equation:

$$I = \sqrt{\frac{P}{R}} = \sqrt{\frac{411\ \text{W}}{0.350\ \Omega}} = 34.3\ \text{A}.$$

Discussion for (b)
The cold current is remarkably higher than the steady-state value of 2.50 A, but the current will quickly decline to that value as the bulb's temperature increases. Most fuses and circuit breakers (used to limit the current in a circuit) are designed to tolerate very high currents briefly as a device comes on. In some cases, such as with electric motors, the current remains high for several seconds, necessitating special "slow blow" fuses.

The Cost of Electricity

The more electric appliances you use and the longer they are left on, the higher your electric bill. This familiar fact is based on the relationship between energy and power. You pay for the energy used. Since $P = E/t$, we see that
Equation:

$$E = \mathrm{Pt}$$

is the energy used by a device using power P for a time interval t. For example, the more lightbulbs burning, the greater P used; the longer they are on, the greater t is. The energy unit on electric bills is the kilowatt-hour $(\mathrm{kW \cdot h})$, consistent with the relationship $E = \mathrm{Pt}$. It is easy to estimate the cost of operating electric appliances if you have some idea of their power consumption rate in watts or kilowatts, the time they are on in hours, and the cost per kilowatt-hour for your electric utility. Kilowatt-hours, like all other specialized energy units such as food calories, can be converted to joules. You can prove to yourself that $1\ \mathrm{kW \cdot h} = 3.6 \times 10^6\ \mathrm{J}$.

The electrical energy (E) used can be reduced either by reducing the time of use or by reducing the power consumption of that appliance or fixture. This will not only reduce the cost, but it will also result in a reduced impact on the environment. Improvements to lighting are some of the fastest ways to reduce the electrical energy used in a home or business. About 20% of a home's use of energy goes to lighting, while the number for commercial establishments is closer to 40%. Fluorescent lights are about four times more efficient than incandescent lights—this is true for both the long tubes and the compact fluorescent lights (CFL). (See [link](b).) Thus, a 60-W incandescent bulb can be replaced by a 15-W CFL, which has the same brightness and color. CFLs have a bent tube inside a globe or a spiral-shaped tube, all connected to a standard screw-in base that fits standard incandescent light sockets. (Original problems with color, flicker, shape, and high initial investment for CFLs have been addressed in recent years.) The heat transfer from these CFLs is less, and they last up to 10 times longer. The significance of an investment in such bulbs is addressed in the next example. New white LED lights (which are clusters of small LED bulbs) are even more efficient (twice that of CFLs) and last 5 times longer than CFLs. However, their cost is still high.

Note:
Making Connections: Energy, Power, and Time

The relationship $E = \mathrm{Pt}$ is one that you will find useful in many different contexts. The energy your body uses in exercise is related to the power level and duration of your activity, for example. The amount of heating by a power source is related to the power level and time it is applied. Even the radiation dose of an X-ray image is related to the power and time of exposure.

Example:
Calculating the Cost Effectiveness of Compact Fluorescent Lights (CFL)

If the cost of electricity in your area is 12 cents per kWh, what is the total cost (capital plus operation) of using a 60-W incandescent bulb for 1000 hours (the lifetime of that bulb) if the bulb cost 25 cents? (b) If we replace this bulb with a compact fluorescent light that provides the same light output, but at one-quarter the wattage, and which costs $1.50 but lasts 10 times longer (10,000 hours), what will that total cost be?

Strategy

To find the operating cost, we first find the energy used in kilowatt-hours and then multiply by the cost per kilowatt-hour.

Solution for (a)

The energy used in kilowatt-hours is found by entering the power and time into the expression for energy:

Equation:

$$E = \mathrm{Pt} = (60\ \mathrm{W})(1000\ \mathrm{h}) = 60{,}000\ \mathrm{W}\cdot\mathrm{h}.$$

In kilowatt-hours, this is

Equation:

$$E = 60.0\ \mathrm{kW}\cdot\mathrm{h}.$$

Now the electricity cost is

Equation:

$$\mathrm{cost} = (60.0\ \mathrm{kW}\cdot\mathrm{h})(\$0.12/\mathrm{kW}\cdot\mathrm{h}) = \$7.20.$$

The total cost will be $7.20 for 1000 hours (about one-half year at 5 hours per day).

Solution for (b)

Since the CFL uses only 15 W and not 60 W, the electricity cost will be $7.20/4 = $1.80. The CFL will last 10 times longer than the incandescent, so that the investment cost will be 1/10 of the bulb cost for that time period of use, or 0.1($1.50) = $0.15. Therefore, the total cost will be $1.95 for 1000 hours.

Discussion

Therefore, it is much cheaper to use the CFLs, even though the initial investment is higher. The increased cost of labor that a business must include for replacing the incandescent bulbs more often has not been figured in here.

Note:

Making Connections: Take-Home Experiment—Electrical Energy Use Inventory

1) Make a list of the power ratings on a range of appliances in your home or room. Explain why something like a toaster has a higher rating than a digital clock. Estimate the energy consumed by these appliances in an average day (by estimating their time of use). Some appliances might only state the operating current. If the household voltage is 120 V, then use $P = \mathrm{IV}$. 2) Check out the total wattage used in the rest rooms of your school's floor or building. (You might need to assume the long fluorescent lights in use are rated at 32 W.) Suppose that the building was closed all weekend and that these lights were left on from 6 p.m. Friday until 8 a.m. Monday. What would this oversight cost? How about for an entire year of weekends?

Section Summary

- Electric power P is the rate (in watts) that energy is supplied by a source or dissipated by a device.

- Three expressions for electrical power are
  Equation:

  $$P = \text{IV},$$

  Equation:

  $$P = \frac{V^2}{R},$$

  and
  Equation:

  $$P = I^2 R.$$

- The energy used by a device with a power P over a time t is $E = \text{Pt}$.

Conceptual Questions

Exercise:

Problem:

Why do incandescent lightbulbs grow dim late in their lives, particularly just before their filaments break?

Exercise:

Problem:

The power dissipated in a resistor is given by $P = V^2/R$, which means power decreases if resistance increases. Yet this power is also given by $P = I^2 R$, which means power increases if resistance increases. Explain why there is no contradiction here.

Problem Exercises

Exercise:

Problem:

What is the power of a 1.00×10^2 MV lightning bolt having a current of 2.00×10^4 A?

Solution:

2.00×10^{12} W

Exercise:

Problem:

What power is supplied to the starter motor of a large truck that draws 250 A of current from a 24.0-V battery hookup?

Exercise:

Problem:

A charge of 4.00 C of charge passes through a pocket calculator's solar cells in 4.00 h. What is the power output, given the calculator's voltage output is 3.00 V? (See [link].)

The strip of solar cells just above the keys of this calculator convert

light to electricity to supply its energy needs. (credit: Evan-Amos, Wikimedia Commons)

Exercise:

Problem:

How many watts does a flashlight that has 6.00×10^2 C pass through it in 0.500 h use if its voltage is 3.00 V?

Exercise:

Problem:

Find the power dissipated in each of these extension cords: (a) an extension cord having a $0.0600 \text{ - } \Omega$ resistance and through which 5.00 A is flowing; (b) a cheaper cord utilizing thinner wire and with a resistance of $0.300\ \Omega$.

Solution:

(a) 1.50 W

(b) 7.50 W

Exercise:

Problem:

Verify that the units of a volt-ampere are watts, as implied by the equation $P = \text{IV}$.

Exercise:

Problem:

Show that the units $1\ \mathrm{V}^2/\Omega = 1\mathrm{W}$, as implied by the equation $P = V^2/R$.

Solution:

$\frac{V^2}{\Omega} = \frac{V^2}{\mathrm{V/A}} = \mathrm{AV} = \left(\frac{C}{s}\right)\left(\frac{J}{C}\right) = \frac{J}{s} = 1\ \mathrm{W}$

Exercise:

Problem:

Show that the units $1\ \mathrm{A}^2 \cdot \Omega = 1\ \mathrm{W}$, as implied by the equation $P = I^2R$.

Exercise:

Problem:

Verify the energy unit equivalence that $1\ \mathrm{kW} \cdot \mathrm{h} = 3.60 \times 10^6\ \mathrm{J}$.

Solution:

$1\ \mathrm{kW} \cdot \mathrm{h} = \left(\frac{1 \times 10^3\ \mathrm{J}}{1\ \mathrm{s}}\right)(1\ \mathrm{h})\left(\frac{3600\ \mathrm{s}}{1\ \mathrm{h}}\right) = 3.60 \times 10^6\ \mathrm{J}$

Exercise:

Problem:

Electrons in an X-ray tube are accelerated through $1.00 \times 10^2\ \mathrm{kV}$ and directed toward a target to produce X-rays. Calculate the power of the electron beam in this tube if it has a current of 15.0 mA.

Exercise:

Problem:

An electric water heater consumes 5.00 kW for 2.00 h per day. What is the cost of running it for one year if electricity costs $12.0\ \text{cents/kW}\cdot\text{h}$? See [link].

On-demand electric hot water heater. Heat is supplied to water only when needed. (credit: aviddavid, Flickr)

Solution:

$438/y

Exercise:

Problem:

With a 1200-W toaster, how much electrical energy is needed to make a slice of toast (cooking time = 1 minute)? At $9.0\ \text{cents/kW}\cdot\text{h}$, how much does this cost?

Exercise:

Problem:

What would be the maximum cost of a CFL such that the total cost (investment plus operating) would be the same for both CFL and incandescent 60-W bulbs? Assume the cost of the incandescent bulb is 25 cents and that electricity costs $10\ \text{cents/kWh}$. Calculate the cost for 1000 hours, as in the cost effectiveness of CFL example.

Solution:

$6.25

Exercise:

Problem:

Some makes of older cars have 6.00-V electrical systems. (a) What is the hot resistance of a 30.0-W headlight in such a car? (b) What current flows through it?

Exercise:

Problem:

Alkaline batteries have the advantage of putting out constant voltage until very nearly the end of their life. How long will an alkaline battery rated at $1.00\ \text{A} \cdot \text{h}$ and 1.58 V keep a 1.00-W flashlight bulb burning?

Solution:

1.58 h

Exercise:

Problem:

A cauterizer, used to stop bleeding in surgery, puts out 2.00 mA at 15.0 kV. (a) What is its power output? (b) What is the resistance of the path?

Exercise:

Problem:

The average television is said to be on 6 hours per day. Estimate the yearly cost of electricity to operate 100 million TVs, assuming their power consumption averages 150 W and the cost of electricity averages $12.0\text{ cents/kW}\cdot\text{h}$.

Solution:

$3.94 billion/year

Exercise:

Problem:

An old lightbulb draws only 50.0 W, rather than its original 60.0 W, due to evaporative thinning of its filament. By what factor is its diameter reduced, assuming uniform thinning along its length? Neglect any effects caused by temperature differences.

Exercise:

Problem:

00-gauge copper wire has a diameter of 9.266 mm. Calculate the power loss in a kilometer of such wire when it carries $1.00\times10^{2}\text{ A}$.

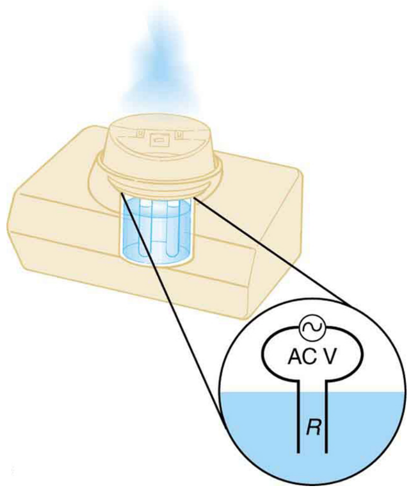
Solution:

25.5 W

Exercise:

Problem: Integrated Concepts

Cold vaporizers pass a current through water, evaporating it with only a small increase in temperature. One such home device is rated at 3.50 A and utilizes 120 V AC with 95.0% efficiency. (a) What is the vaporization rate in grams per minute? (b) How much water must you put into the vaporizer for 8.00 h of overnight operation? (See [link].)



This cold vaporizer passes current directly through water, vaporizing it directly with relatively little temperature increase.

Exercise:

Problem: Integrated Concepts

(a) What energy is dissipated by a lightning bolt having a 20,000-A current, a voltage of 1.00×10^2 MV, and a length of 1.00 ms? (b) What mass of tree sap could be raised from 18.0°C to its boiling point and then evaporated by this energy, assuming sap has the same thermal characteristics as water?

Solution:

(a) 2.00×10^9 J

(b) 769 kg

Exercise:

Problem: Integrated Concepts

What current must be produced by a 12.0-V battery-operated bottle warmer in order to heat 75.0 g of glass, 250 g of baby formula, and 3.00×10^2 g of aluminum from $20.0^{\circ}\mathrm{C}$ to $90.0^{\circ}\mathrm{C}$ in 5.00 min?

Exercise:

Problem: Integrated Concepts

How much time is needed for a surgical cauterizer to raise the temperature of 1.00 g of tissue from $37.0^{\circ}\mathrm{C}$ to $100^{\circ}\mathrm{C}$ and then boil away 0.500 g of water, if it puts out 2.00 mA at 15.0 kV? Ignore heat transfer to the surroundings.

Solution:

45.0 s

Exercise:

Problem: Integrated Concepts

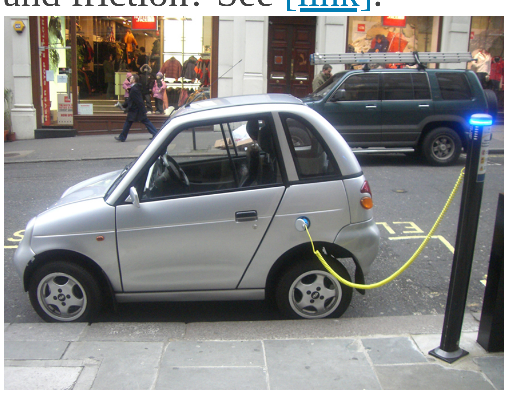
Hydroelectric generators (see [link]) at Hoover Dam produce a maximum current of 8.00×10^3 A at 250 kV. (a) What is the power output? (b) The water that powers the generators enters and leaves the system at low speed (thus its kinetic energy does not change) but loses 160 m in altitude. How many cubic meters per second are needed, assuming 85.0% efficiency?

Hydroelectric generators at the Hoover dam. (credit: Jon Sullivan)

Exercise:

Problem: Integrated Concepts

(a) Assuming 95.0% efficiency for the conversion of electrical power by the motor, what current must the 12.0-V batteries of a 750-kg electric car be able to supply: (a) To accelerate from rest to 25.0 m/s in 1.00 min? (b) To climb a $2.00 \times 10^2\text{-m}$-high hill in 2.00 min at a constant 25.0-m/s speed while exerting $5.00 \times 10^2\ \text{N}$ of force to overcome air resistance and friction? (c) To travel at a constant 25.0-m/s speed, exerting a $5.00 \times 10^2\ \text{N}$ force to overcome air resistance and friction? See [link].



This REVAi, an electric

car, gets recharged on a street in London. (credit: Frank Hebbert)

Solution:

(a) 343 A

(b) 2.17×10^3 A

(c) 1.10×10^3 A

Exercise:

Problem: Integrated Concepts

A light-rail commuter train draws 630 A of 650-V DC electricity when accelerating. (a) What is its power consumption rate in kilowatts? (b) How long does it take to reach 20.0 m/s starting from rest if its loaded mass is 5.30×10^4 kg, assuming 95.0% efficiency and constant power? (c) Find its average acceleration. (d) Discuss how the acceleration you found for the light-rail train compares to what might be typical for an automobile.

Exercise:

Problem: Integrated Concepts

(a) An aluminum power transmission line has a resistance of $0.0580\ \Omega/\text{km}$. What is its mass per kilometer? (b) What is the mass per kilometer of a copper line having the same resistance? A lower resistance would shorten the heating time. Discuss the practical limits to speeding the heating by lowering the resistance.

Solution:

(a) 1.23×10^3 kg

(b) 2.64×10^3 kg

Exercise:

Problem: Integrated Concepts

(a) An immersion heater utilizing 120 V can raise the temperature of a 1.00×10^2-g aluminum cup containing 350 g of water from 20.0°C to 95.0°C in 2.00 min. Find its resistance, assuming it is constant during the process. (b) A lower resistance would shorten the heating time. Discuss the practical limits to speeding the heating by lowering the resistance.

Exercise:

Problem: Integrated Concepts

(a) What is the cost of heating a hot tub containing 1500 kg of water from 10.0°C to 40.0°C, assuming 75.0% efficiency to account for heat transfer to the surroundings? The cost of electricity is $9 \text{ cents/kW} \cdot \text{h}$. (b) What current was used by the 220-V AC electric heater, if this took 4.00 h?

Exercise:

Problem: Unreasonable Results

(a) What current is needed to transmit 1.00×10^2 MW of power at 480 V? (b) What power is dissipated by the transmission lines if they have a $1.00 \text{ - } \Omega$ resistance? (c) What is unreasonable about this result? (d) Which assumptions are unreasonable, or which premises are inconsistent?

Solution:

(a) 2.08×10^5 A

(b) 4.33×10^4 MW

(c) The transmission lines dissipate more power than they are supposed to transmit.

(d) A voltage of 480 V is unreasonably low for a transmission voltage. Long-distance transmission lines are kept at much higher voltages (often hundreds of kilovolts) to reduce power losses.

Exercise:

Problem: Unreasonable Results

(a) What current is needed to transmit 1.00×10^2 MW of power at 10.0 kV? (b) Find the resistance of 1.00 km of wire that would cause a 0.0100% power loss. (c) What is the diameter of a 1.00-km-long copper wire having this resistance? (d) What is unreasonable about these results? (e) Which assumptions are unreasonable, or which premises are inconsistent?

Exercise:

Problem: Construct Your Own Problem

Consider an electric immersion heater used to heat a cup of water to make tea. Construct a problem in which you calculate the needed resistance of the heater so that it increases the temperature of the water and cup in a reasonable amount of time. Also calculate the cost of the electrical energy used in your process. Among the things to be considered are the voltage used, the masses and heat capacities involved, heat losses, and the time over which the heating takes place. Your instructor may wish for you to consider a thermal safety switch (perhaps bimetallic) that will halt the process before damaging temperatures are reached in the immersion unit.

Glossary

electric power

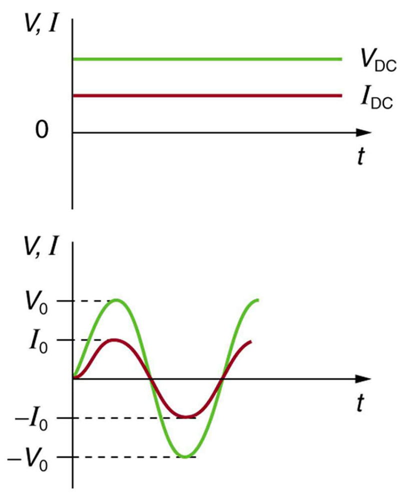
the rate at which electrical energy is supplied by a source or dissipated by a device; it is the product of current times voltage

Alternating Current versus Direct Current

- Explain the differences and similarities between AC and DC current.
- Calculate rms voltage, current, and average power.
- Explain why AC current is used for power transmission.

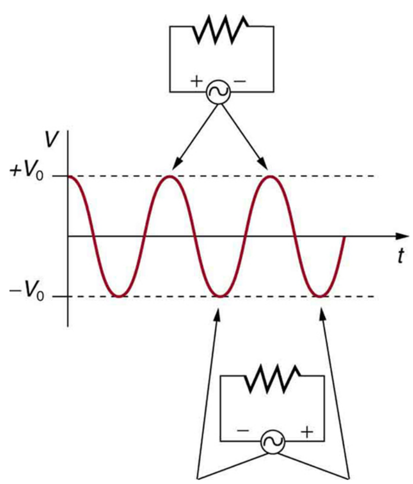
Alternating Current

Most of the examples dealt with so far, and particularly those utilizing batteries, have constant voltage sources. Once the current is established, it is thus also a constant. **Direct current** (DC) is the flow of electric charge in only one direction. It is the steady state of a constant-voltage circuit. Most well-known applications, however, use a time-varying voltage source. **Alternating current** (AC) is the flow of electric charge that periodically reverses direction. If the source varies periodically, particularly sinusoidally, the circuit is known as an alternating current circuit. Examples include the commercial and residential power that serves so many of our needs. [link] shows graphs of voltage and current versus time for typical DC and AC power. The AC voltages and frequencies commonly used in homes and businesses vary around the world.



(a) DC voltage and current are constant in time, once the

current is established. (b) A graph of voltage and current versus time for 60-Hz AC power. The voltage and current are sinusoidal and are in phase for a simple resistance circuit. The frequencies and peak voltages of AC sources differ greatly.



The potential difference V between the terminals of an AC voltage source fluctuates as

shown. The mathematical expression for V is given by $V = V_0 \sin 2\pi \text{ft}$.

[link] shows a schematic of a simple circuit with an AC voltage source. The voltage between the terminals fluctuates as shown, with the **AC voltage** given by

Equation:

$$V = V_0 \text{ sin } 2\pi \text{ft},$$

where V is the voltage at time t, V_0 is the peak voltage, and f is the frequency in hertz. For this simple resistance circuit, $I = \text{V/R}$, and so the **AC current** is

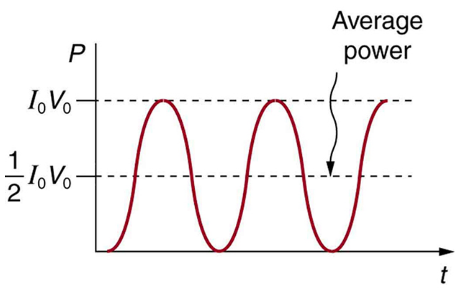
Equation:

$$I = I_0 \text{ sin } 2\pi \text{ft},$$

where I is the current at time t, and $I_0 = V_0/\text{R}$ is the peak current. For this example, the voltage and current are said to be in phase, as seen in [link](b).

Current in the resistor alternates back and forth just like the driving voltage, since $I = \text{V/R}$. If the resistor is a fluorescent light bulb, for example, it brightens and dims 120 times per second as the current repeatedly goes through zero. A 120-Hz flicker is too rapid for your eyes to detect, but if you wave your hand back and forth between your face and a fluorescent light, you will see a stroboscopic effect evidencing AC. The fact that the light output fluctuates means that the power is fluctuating. The power supplied is $P = \text{IV}$. Using the expressions for I and V above, we see that the time dependence of power is $P = I_0 V_0 \text{ sin}^2 \text{ } 2\pi \text{ft}$, as shown in [link].

Note:
Making Connections: Take-Home Experiment—AC/DC Lights
Wave your hand back and forth between your face and a fluorescent light bulb. Do you observe the same thing with the headlights on your car? Explain what you observe. *Warning: Do not look directly at very bright light*.



AC power as a function of time. Since the voltage and current are in phase here, their product is non-negative and fluctuates between zero and I_0V_0. Average power is $(1/2)I_0V_0$.

We are most often concerned with average power rather than its fluctuations—that 60-W light bulb in your desk lamp has an average power consumption of 60 W, for example. As illustrated in [link], the average power P_{ave} is

Equation:

$$P_{\text{ave}} = \frac{1}{2}I_0V_0.$$

This is evident from the graph, since the areas above and below the $(1/2)I_0V_0$ line are equal, but it can also be proven using trigonometric identities. Similarly, we define an average or **rms current** I_{rms} and average or **rms voltage** V_{rms} to be, respectively,

Equation:

$$I_{\text{rms}} = \frac{I_0}{\sqrt{2}}$$

and

Equation:

$$V_{\text{rms}} = \frac{V_0}{\sqrt{2}}.$$

where rms stands for root mean square, a particular kind of average. In general, to obtain a root mean square, the particular quantity is squared, its mean (or average) is found, and the square root is taken. This is useful for AC, since the average value is zero. Now,

Equation:

$$P_{\text{ave}} = I_{\text{rms}}V_{\text{rms}},$$

which gives

Equation:

$$P_{\text{ave}} = \frac{I_0}{\sqrt{2}} \cdot \frac{V_0}{\sqrt{2}} = \frac{1}{2}I_0V_0,$$

as stated above. It is standard practice to quote I_{rms}, V_{rms}, and P_{ave} rather than the peak values. For example, most household electricity is 120 V AC, which means that V_{rms} is 120 V. The common 10-A circuit breaker will interrupt a sustained I_{rms} greater than 10 A. Your 1.0-kW microwave oven

consumes $P_{\text{ave}} = 1.0\text{ kW}$, and so on. You can think of these rms and average values as the equivalent DC values for a simple resistive circuit.

To summarize, when dealing with AC, Ohm's law and the equations for power are completely analogous to those for DC, but rms and average values are used for AC. Thus, for AC, Ohm's law is written

Equation:

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{R}.$$

The various expressions for AC power P_{ave} are

Equation:

$$P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}},$$

Equation:

$$P_{\text{ave}} = \frac{V_{\text{rms}}^2}{R},$$

and

Equation:

$$P_{\text{ave}} = I_{\text{rms}}^2 R.$$

Example:

Peak Voltage and Power for AC

(a) What is the value of the peak voltage for 120-V AC power? (b) What is the peak power consumption rate of a 60.0-W AC light bulb?

Strategy

We are told that V_{rms} is 120 V and P_{ave} is 60.0 W. We can use $V_{\text{rms}} = \frac{V_0}{\sqrt{2}}$ to find the peak voltage, and we can manipulate the definition of power to

find the peak power from the given average power.

Solution for (a)

Solving the equation $V_{\text{rms}} = \frac{V_0}{\sqrt{2}}$ for the peak voltage V_0 and substituting the known value for V_{rms} gives

Equation:

$$V_0 = \sqrt{2}V_{\text{rms}} = 1.414(120\text{ V}) = 170\text{ V}.$$

Discussion for (a)

This means that the AC voltage swings from 170 V to –170 V and back 60 times every second. An equivalent DC voltage is a constant 120 V.

Solution for (b)

Peak power is peak current times peak voltage. Thus,

Equation:

$$P_0 = I_0V_0 = 2\left(\frac{1}{2}I_0V_0\right) = 2P_{\text{ave}}.$$

We know the average power is 60.0 W, and so

Equation:

$$P_0 = 2(60.0\text{ W}) = 120\text{ W}.$$

Discussion

So the power swings from zero to 120 W one hundred twenty times per second (twice each cycle), and the power averages 60 W.

Why Use AC for Power Distribution?

Most large power-distribution systems are AC. Moreover, the power is transmitted at much higher voltages than the 120-V AC (240 V in most parts of the world) we use in homes and on the job. Economies of scale make it cheaper to build a few very large electric power-generation plants than to build numerous small ones. This necessitates sending power long distances, and it is obviously important that energy losses en route be

minimized. High voltages can be transmitted with much smaller power losses than low voltages, as we shall see. (See [link].) For safety reasons, the voltage at the user is reduced to familiar values. The crucial factor is that it is much easier to increase and decrease AC voltages than DC, so AC is used in most large power distribution systems.

Power is distributed over large distances at high voltage to reduce power loss in the transmission lines. The voltages generated at the power plant are stepped up by passive devices called transformers (see Transformers) to 330,000 volts (or more in some places worldwide). At the point of use, the transformers reduce the voltage transmitted for safe residential and commercial use. (Credit: GeorgHH, Wikimedia Commons)

Example:

Power Losses Are Less for High-Voltage Transmission

(a) What current is needed to transmit 100 MW of power at 200 kV? (b) What is the power dissipated by the transmission lines if they have a resistance of 1.00 Ω? (c) What percentage of the power is lost in the transmission lines?

Strategy

We are given $P_{\text{ave}} = 100\text{ MW}$, $V_{\text{rms}} = 200\text{ kV}$, and the resistance of the lines is $R = 1.00\ \Omega$. Using these givens, we can find the current flowing (from $P = \text{IV}$) and then the power dissipated in the lines ($P = I^2R$), and we take the ratio to the total power transmitted.

Solution

To find the current, we rearrange the relationship $P_{\text{ave}} = I_{\text{rms}}V_{\text{rms}}$ and substitute known values. This gives

Equation:

$$I_{\text{rms}} = \frac{P_{\text{ave}}}{V_{\text{rms}}} = \frac{100 \times 10^6\text{ W}}{200 \times 10^3\text{ V}} = 500\text{ A}.$$

Solution

Knowing the current and given the resistance of the lines, the power dissipated in them is found from $P_{\text{ave}} = I_{\text{rms}}^2 R$. Substituting the known values gives

Equation:

$$P_{\text{ave}} = I_{\text{rms}}^2 R = (500\text{ A})^2(1.00\ \Omega) = 250\text{ kW}.$$

Solution

The percent loss is the ratio of this lost power to the total or input power, multiplied by 100:

Equation:

$$\% \text{ loss} = \frac{250 \text{ kW}}{100 \text{ MW}} \times 100 = 0.250 \text{ \%}.$$

Discussion

One-fourth of a percent is an acceptable loss. Note that if 100 MW of power had been transmitted at 25 kV, then a current of 4000 A would have been needed. This would result in a power loss in the lines of 16.0 MW, or 16.0% rather than 0.250%. The lower the voltage, the more current is needed, and the greater the power loss in the fixed-resistance transmission lines. Of course, lower-resistance lines can be built, but this requires larger and more expensive wires. If superconducting lines could be economically produced, there would be no loss in the transmission lines at all. But, as we shall see in a later chapter, there is a limit to current in superconductors, too. In short, high voltages are more economical for transmitting power, and AC voltage is much easier to raise and lower, so that AC is used in most large-scale power distribution systems.

It is widely recognized that high voltages pose greater hazards than low voltages. But, in fact, some high voltages, such as those associated with common static electricity, can be harmless. So it is not voltage alone that determines a hazard. It is not so widely recognized that AC shocks are often more harmful than similar DC shocks. Thomas Edison thought that AC shocks were more harmful and set up a DC power-distribution system in New York City in the late 1800s. There were bitter fights, in particular between Edison and George Westinghouse and Nikola Tesla, who were advocating the use of AC in early power-distribution systems. AC has prevailed largely due to transformers and lower power losses with high-voltage transmission.

Note:

PhET Explorations: Generator

Generate electricity with a bar magnet! Discover the physics behind the phenomena by exploring magnets and how you can use them to make a bulb light.

[Generator]

Section Summary

- Direct current (DC) is the flow of electric current in only one direction. It refers to systems where the source voltage is constant.
- The voltage source of an alternating current (AC) system puts out $V = V_0 \sin 2\pi \text{ft}$, where V is the voltage at time t, V_0 is the peak voltage, and f is the frequency in hertz.
- In a simple circuit, $I = \text{V}/\text{R}$ and AC current is $I = I_0 \sin 2\pi \text{ft}$, where I is the current at time t, and $I_0 = V_0/\text{R}$ is the peak current.
- The average AC power is $P_{\text{ave}} = \frac{1}{2} I_0 V_0$.
- Average (rms) current I_{rms} and average (rms) voltage V_{rms} are $I_{\text{rms}} = \frac{I_0}{\sqrt{2}}$ and $V_{\text{rms}} = \frac{V_0}{\sqrt{2}}$, where rms stands for root mean square.
- Thus, $P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}}$.
- Ohm's law for AC is $I_{\text{rms}} = \frac{V_{\text{rms}}}{R}$.
- Expressions for the average power of an AC circuit are $P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}}$, $P_{\text{ave}} = \frac{{V_{\text{rms}}}^2}{R}$, and $P_{\text{ave}} = {I_{\text{rms}}}^2 R$, analogous to the expressions for DC circuits.

Conceptual Questions

Exercise:

Problem:

Give an example of a use of AC power other than in the household. Similarly, give an example of a use of DC power other than that supplied by batteries.

Exercise:

Problem:

Why do voltage, current, and power go through zero 120 times per second for 60-Hz AC electricity?

Exercise:

Problem:

You are riding in a train, gazing into the distance through its window. As close objects streak by, you notice that the nearby fluorescent lights make *dashed* streaks. Explain.

Problem Exercises

Exercise:

Problem:

(a) What is the hot resistance of a 25-W light bulb that runs on 120-V AC? (b) If the bulb's operating temperature is $2700^{\circ}\mathrm{C}$, what is its resistance at $2600^{\circ}\mathrm{C}$?

Exercise:

Problem:

Certain heavy industrial equipment uses AC power that has a peak voltage of 679 V. What is the rms voltage?

Solution:

480 V

Exercise:

Problem:

A certain circuit breaker trips when the rms current is 15.0 A. What is the corresponding peak current?

Exercise:

Problem:

Military aircraft use 400-Hz AC power, because it is possible to design lighter-weight equipment at this higher frequency. What is the time for one complete cycle of this power?

Solution:

2.50 ms

Exercise:

Problem:

A North American tourist takes his 25.0-W, 120-V AC razor to Europe, finds a special adapter, and plugs it into 240 V AC. Assuming constant resistance, what power does the razor consume as it is ruined?

Exercise:

Problem:

In this problem, you will verify statements made at the end of the power losses for [link]. (a) What current is needed to transmit 100 MW of power at a voltage of 25.0 kV? (b) Find the power loss in a 1.00 - Ω transmission line. (c) What percent loss does this represent?

Solution:

(a) 4.00 kA

(b) 16.0 MW

(c) 16.0%

Exercise:

Problem:

A small office-building air conditioner operates on 408-V AC and consumes 50.0 kW. (a) What is its effective resistance? (b) What is the cost of running the air conditioner during a hot summer month when it is on 8.00 h per day for 30 days and electricity costs $9.00\ \text{cents/kW} \cdot \text{h}$?

Exercise:

Problem:

What is the peak power consumption of a 120-V AC microwave oven that draws 10.0 A?

Solution:

2.40 kW

Exercise:

Problem:

What is the peak current through a 500-W room heater that operates on 120-V AC power?

Exercise:

Problem:

Two different electrical devices have the same power consumption, but one is meant to be operated on 120-V AC and the other on 240-V AC. (a) What is the ratio of their resistances? (b) What is the ratio of their currents? (c) Assuming its resistance is unaffected, by what factor will the power increase if a 120-V AC device is connected to 240-V AC?

Solution:

(a) 4.0

(b) 0.50

(c) 4.0

Exercise:

Problem:

Nichrome wire is used in some radiative heaters. (a) Find the resistance needed if the average power output is to be 1.00 kW utilizing 120-V AC. (b) What length of Nichrome wire, having a cross-sectional area of 5.00mm^2, is needed if the operating temperature is 500^{o} C? (c) What power will it draw when first switched on?

Exercise:

Problem:

Find the time after $t = 0$ when the instantaneous voltage of 60-Hz AC first reaches the following values: (a) $V_0/2$ (b) V_0 (c) 0.

Solution:

(a) 1.39 ms

(b) 4.17 ms

(c) 8.33 ms

Exercise:

Problem:

(a) At what two times in the first period following $t = 0$ does the instantaneous voltage in 60-Hz AC equal V_{rms}? (b) $-V_{\text{rms}}$?

Glossary

direct current
: (DC) the flow of electric charge in only one direction

alternating current
: (AC) the flow of electric charge that periodically reverses direction

AC voltage

voltage that fluctuates sinusoidally with time, expressed as $V = V_0 \sin 2\pi ft$, where V is the voltage at time t, V_0 is the peak voltage, and f is the frequency in hertz

AC current

current that fluctuates sinusoidally with time, expressed as $I = I_0 \sin 2\pi ft$, where I is the current at time t, I_0 is the peak current, and f is the frequency in hertz

rms current

the root mean square of the current, $I_{\text{rms}} = I_0/\sqrt{2}$, where I_0 is the peak current, in an AC system

rms voltage

the root mean square of the voltage, $V_{\text{rms}} = V_0/\sqrt{2}$, where V_0 is the peak voltage, in an AC system

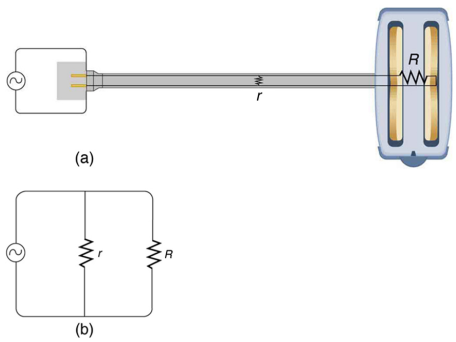
Electric Hazards and the Human Body

- Define thermal hazard, shock hazard, and short circuit.
- Explain what effects various levels of current have on the human body.

There are two known hazards of electricity—thermal and shock. A **thermal hazard** is one where excessive electric power causes undesired thermal effects, such as starting a fire in the wall of a house. A **shock hazard** occurs when electric current passes through a person. Shocks range in severity from painful, but otherwise harmless, to heart-stopping lethality. This section considers these hazards and the various factors affecting them in a quantitative manner. Electrical Safety: Systems and Devices will consider systems and devices for preventing electrical hazards.

Thermal Hazards

Electric power causes undesired heating effects whenever electric energy is converted to thermal energy at a rate faster than it can be safely dissipated. A classic example of this is the **short circuit**, a low-resistance path between terminals of a voltage source. An example of a short circuit is shown in [link]. Insulation on wires leading to an appliance has worn through, allowing the two wires to come into contact. Such an undesired contact with a high voltage is called a *short*. Since the resistance of the short, r, is very small, the power dissipated in the short, $P=V^2/r$, is very large. For example, if V is 120 V and r is $0.100\ \Omega$, then the power is 144 kW, *much* greater than that used by a typical household appliance. Thermal energy delivered at this rate will very quickly raise the temperature of surrounding materials, melting or perhaps igniting them.

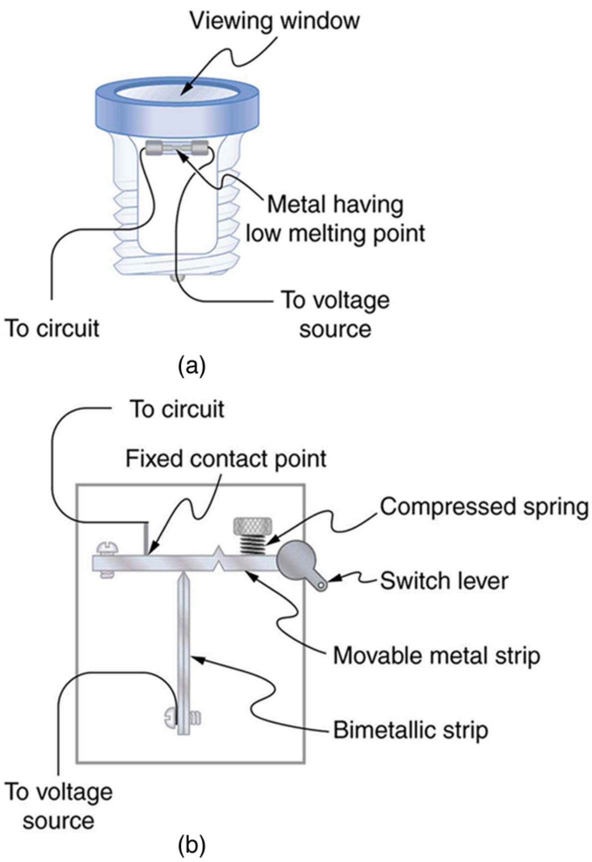


A short circuit is an undesired low-resistance path across a voltage source. (a) Worn insulation on the wires of a toaster allow them to come into contact with a low resistance r. Since $P = V^2/r$, thermal power is created so rapidly that the cord melts or burns. (b) A schematic of the short circuit.

One particularly insidious aspect of a short circuit is that its resistance may actually be decreased due to the increase in temperature. This can happen if the short creates ionization. These charged atoms and molecules are free to move and, thus, lower the resistance r. Since $P = V^2/r$, the power dissipated in the short rises, possibly causing more ionization, more power, and so on. High voltages, such as the 480-V AC used in some industrial applications, lend themselves to this hazard, because higher voltages create higher initial power production in a short.

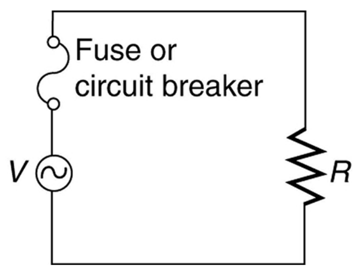
Another serious, but less dramatic, thermal hazard occurs when wires supplying power to a user are overloaded with too great a current. As

discussed in the previous section, the power dissipated in the supply wires is $P = I^2 R_{\mathrm{w}}$, where R_{w} is the resistance of the wires and I the current flowing through them. If either I or R_{w} is too large, the wires overheat. For example, a worn appliance cord (with some of its braided wires broken) may have $R_{\mathrm{w}} = 2.00 \ \Omega$ rather than the $0.100 \ \Omega$ it should be. If 10.0 A of current passes through the cord, then $P = I^2 R_{\mathrm{w}} = 200 \ \mathrm{W}$ is dissipated in the cord—much more than is safe. Similarly, if a wire with a 0.100 - Ω resistance is meant to carry a few amps, but is instead carrying 100 A, it will severely overheat. The power dissipated in the wire will in that case be $P = 1000 \ \mathrm{W}$. Fuses and circuit breakers are used to limit excessive currents. (See [link] and [link].) Each device opens the circuit automatically when a sustained current exceeds safe limits.



(a) A fuse has a metal strip with a low melting point that, when overheated by an excessive

current, permanently breaks the connection of a circuit to a voltage source. (b) A circuit breaker is an automatic but restorable electric switch. The one shown here has a bimetallic strip that bends to the right and into the notch if overheated. The spring then forces the metal strip downward, breaking the electrical connection at the points.



Schematic of a circuit with a fuse or circuit breaker in it. Fuses and circuit breakers act like automatic switches that open when sustained current exceeds desired limits.

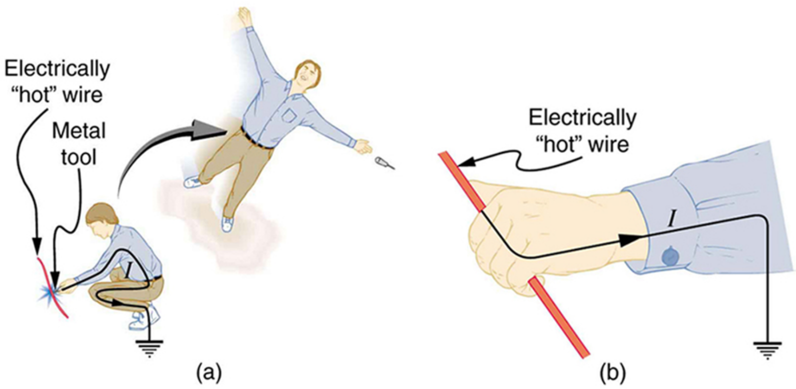
Fuses and circuit breakers for typical household voltages and currents are relatively simple to produce, but those for large voltages and currents experience special problems. For example, when a circuit breaker tries to interrupt the flow of high-voltage electricity, a spark can jump across its points that ionizes the air in the gap and allows the current to continue flowing. Large circuit breakers found in power-distribution systems employ insulating gas and even use jets of gas to blow out such sparks. Here AC is safer than DC, since AC current goes through zero 120 times per second, giving a quick opportunity to extinguish these arcs.

Shock Hazards

Electrical currents through people produce tremendously varied effects. An electrical current can be used to block back pain. The possibility of using electrical current to stimulate muscle action in paralyzed limbs, perhaps allowing paraplegics to walk, is under study. TV dramatizations in which electrical shocks are used to bring a heart attack victim out of ventricular fibrillation (a massively irregular, often fatal, beating of the heart) are more than common. Yet most electrical shock fatalities occur because a current put the heart into fibrillation. A pacemaker uses electrical shocks to stimulate the heart to beat properly. Some fatal shocks do not produce burns, but warts can be safely burned off with electric current (though freezing using liquid nitrogen is now more common). Of course, there are consistent explanations for these disparate effects. The major factors upon which the effects of electrical shock depend are

1. The amount of current I
2. The path taken by the current
3. The duration of the shock
4. The frequency f of the current ($f=0$ for DC)

[link] gives the effects of electrical shocks as a function of current for a typical accidental shock. The effects are for a shock that passes through the trunk of the body, has a duration of 1 s, and is caused by 60-Hz power.



An electric current can cause muscular contractions with varying effects. (a) The victim is "thrown" backward by involuntary muscle contractions that extend the legs and torso. (b) The victim can't let go of the wire that is stimulating all the muscles in the hand. Those that close the fingers are stronger than those that open them.

Current (mA)	**Effect**
1	Threshold of sensation
5	Maximum harmless current
10–20	Onset of sustained muscular contraction; cannot let go for duration of shock; contraction of chest muscles may stop breathing during shock

Current (mA)	Effect
50	Onset of pain
100–300+	Ventricular fibrillation possible; often fatal
300	Onset of burns depending on concentration of current
6000 (6 A)	Onset of sustained ventricular contraction and respiratory paralysis; both cease when shock ends; heartbeat may return to normal; used to defibrillate the heart

Effects of Electrical Shock as a Function of Current[footnote]
For an average male shocked through trunk of body for 1 s by 60-Hz AC. Values for females are 60–80% of those listed.

Our bodies are relatively good conductors due to the water in our bodies. Given that larger currents will flow through sections with lower resistance (to be further discussed in the next chapter), electric currents preferentially flow through paths in the human body that have a minimum resistance in a direct path to earth. The earth is a natural electron sink. Wearing insulating shoes, a requirement in many professions, prohibits a pathway for electrons by providing a large resistance in that path. Whenever working with high-power tools (drills), or in risky situations, ensure that you do not provide a pathway for current flow (especially through the heart).

Very small currents pass harmlessly and unfelt through the body. This happens to you regularly without your knowledge. The threshold of sensation is only 1 mA and, although unpleasant, shocks are apparently harmless for currents less than 5 mA. A great number of safety rules take the 5-mA value for the maximum allowed shock. At 10 to 20 mA and above, the current can stimulate sustained muscular contractions much as regular nerve impulses do. People sometimes say they were knocked across the room by a shock, but what really happened was that certain muscles

contracted, propelling them in a manner not of their own choosing. (See [link](a).) More frightening, and potentially more dangerous, is the “can’t let go” effect illustrated in [link](b). The muscles that close the fingers are stronger than those that open them, so the hand closes involuntarily on the wire shocking it. This can prolong the shock indefinitely. It can also be a danger to a person trying to rescue the victim, because the rescuer’s hand may close about the victim’s wrist. Usually the best way to help the victim is to give the fist a hard knock/blow/jar with an insulator or to throw an insulator at the fist. Modern electric fences, used in animal enclosures, are now pulsed on and off to allow people who touch them to get free, rendering them less lethal than in the past.

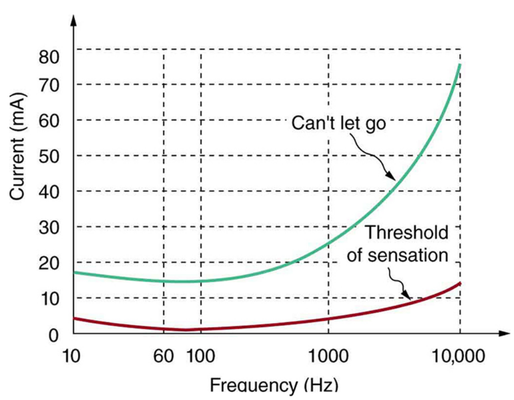
Greater currents may affect the heart. Its electrical patterns can be disrupted, so that it beats irregularly and ineffectively in a condition called “ventricular fibrillation.” This condition often lingers after the shock and is fatal due to a lack of blood circulation. The threshold for ventricular fibrillation is between 100 and 300 mA. At about 300 mA and above, the shock can cause burns, depending on the concentration of current—the more concentrated, the greater the likelihood of burns.

Very large currents cause the heart and diaphragm to contract for the duration of the shock. Both the heart and breathing stop. Interestingly, both often return to normal following the shock. The electrical patterns on the heart are completely erased in a manner that the heart can start afresh with normal beating, as opposed to the permanent disruption caused by smaller currents that can put the heart into ventricular fibrillation. The latter is something like scribbling on a blackboard, whereas the former completely erases it. TV dramatizations of electric shock used to bring a heart attack victim out of ventricular fibrillation also show large paddles. These are used to spread out current passed through the victim to reduce the likelihood of burns.

Current is the major factor determining shock severity (given that other conditions such as path, duration, and frequency are fixed, such as in the table and preceding discussion). A larger voltage is more hazardous, but since $I = \mathrm{V}/\mathrm{R}$, the severity of the shock depends on the combination of voltage and resistance. For example, a person with dry skin has a resistance

of about $200\ \text{k}\Omega$. If he comes into contact with 120-V AC, a current $I = (120\ \text{V})/(200\ \text{k}\Omega) = 0.6\ \text{mA}$ passes harmlessly through him. The same person soaking wet may have a resistance of $10.0\ \text{k}\Omega$ and the same 120 V will produce a current of 12 mA—above the “can’t let go” threshold and potentially dangerous.

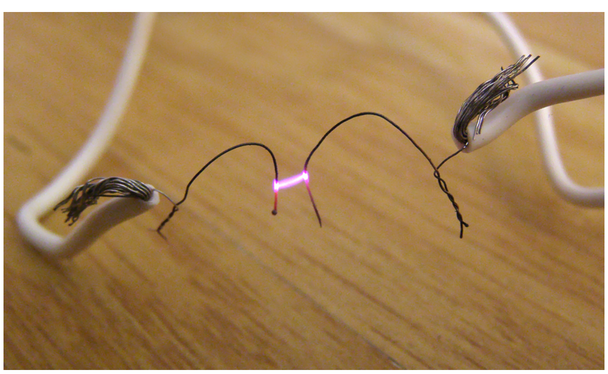
Most of the body’s resistance is in its dry skin. When wet, salts go into ion form, lowering the resistance significantly. The interior of the body has a much lower resistance than dry skin because of all the ionic solutions and fluids it contains. If skin resistance is bypassed, such as by an intravenous infusion, a catheter, or exposed pacemaker leads, a person is rendered **microshock sensitive**. In this condition, currents about 1/1000 those listed in [link] produce similar effects. During open-heart surgery, currents as small as $20\ \mu\text{A}$ can be used to still the heart. Stringent electrical safety requirements in hospitals, particularly in surgery and intensive care, are related to the doubly disadvantaged microshock-sensitive patient. The break in the skin has reduced his resistance, and so the same voltage causes a greater current, and a much smaller current has a greater effect.



Graph of average values for the threshold of sensation and the “can’t let go” current as a function of frequency. The lower the value, the

more sensitive the body is at that frequency.

Factors other than current that affect the severity of a shock are its path, duration, and AC frequency. Path has obvious consequences. For example, the heart is unaffected by an electric shock through the brain, such as may be used to treat manic depression. And it is a general truth that the longer the duration of a shock, the greater its effects. [link] presents a graph that illustrates the effects of frequency on a shock. The curves show the minimum current for two different effects, as a function of frequency. The lower the current needed, the more sensitive the body is at that frequency. Ironically, the body is most sensitive to frequencies near the 50- or 60-Hz frequencies in common use. The body is slightly less sensitive for DC ($f=0$), mildly confirming Edison's claims that AC presents a greater hazard. At higher and higher frequencies, the body becomes progressively less sensitive to any effects that involve nerves. This is related to the maximum rates at which nerves can fire or be stimulated. At very high frequencies, electrical current travels only on the surface of a person. Thus a wart can be burned off with very high frequency current without causing the heart to stop. (Do not try this at home with 60-Hz AC!) Some of the spectacular demonstrations of electricity, in which high-voltage arcs are passed through the air and over people's bodies, employ high frequencies and low currents. (See [link].) Electrical safety devices and techniques are discussed in detail in Electrical Safety: Systems and Devices.



Is this electric arc dangerous?

The answer depends on the AC frequency and the power involved. (credit: Khimich Alex, Wikimedia Commons)

Section Summary

- The two types of electric hazards are thermal (excessive power) and shock (current through a person).
- Shock severity is determined by current, path, duration, and AC frequency.
- [link] lists shock hazards as a function of current.
- [link] graphs the threshold current for two hazards as a function of frequency.

Conceptual Questions

Exercise:

Problem:

Using an ohmmeter, a student measures the resistance between various points on his body. He finds that the resistance between two points on the same finger is about the same as the resistance between two points on opposite hands—both are several hundred thousand ohms. Furthermore, the resistance decreases when more skin is brought into contact with the probes of the ohmmeter. Finally, there is a dramatic drop in resistance (to a few thousand ohms) when the skin is wet. Explain these observations and their implications regarding skin and internal resistance of the human body.

Exercise:

Problem: What are the two major hazards of electricity?

Exercise:

Problem: Why isn't a short circuit a shock hazard?

Exercise:

Problem:

What determines the severity of a shock? Can you say that a certain voltage is hazardous without further information?

Exercise:

Problem:

An electrified needle is used to burn off warts, with the circuit being completed by having the patient sit on a large butt plate. Why is this plate large?

Exercise:

Problem:

Some surgery is performed with high-voltage electricity passing from a metal scalpel through the tissue being cut. Considering the nature of electric fields at the surface of conductors, why would you expect most of the current to flow from the sharp edge of the scalpel? Do you think high- or low-frequency AC is used?

Exercise:

Problem:

Some devices often used in bathrooms, such as hairdryers, often have safety messages saying "Do not use when the bathtub or basin is full of water." Why is this so?

Exercise:

Problem:

We are often advised to not flick electric switches with wet hands, dry your hand first. We are also advised to never throw water on an electric fire. Why is this so?

Exercise:

Problem:

Before working on a power transmission line, linemen will touch the line with the back of the hand as a final check that the voltage is zero. Why the back of the hand?

Exercise:

Problem:

Why is the resistance of wet skin so much smaller than dry, and why do blood and other bodily fluids have low resistances?

Exercise:

Problem:

Could a person on intravenous infusion (an IV) be microshock sensitive?

Exercise:

Problem:

In view of the small currents that cause shock hazards and the larger currents that circuit breakers and fuses interrupt, how do they play a role in preventing shock hazards?

Problem Exercises

Exercise:

Problem:

(a) How much power is dissipated in a short circuit of 240-V AC through a resistance of $0.250\ \Omega$? (b) What current flows?

Solution:

(a) 230 kW

(b) 960 A

Exercise:

Problem:

What voltage is involved in a 1.44-kW short circuit through a 0.100 - Ω resistance?

Exercise:

Problem:

Find the current through a person and identify the likely effect on her if she touches a 120-V AC source: (a) if she is standing on a rubber mat and offers a total resistance of $300\ \text{k}\Omega$; (b) if she is standing barefoot on wet grass and has a resistance of only $4000\ \text{k}\Omega$.

Solution:

(a) 0.400 mA, no effect

(b) 26.7 mA, muscular contraction for duration of the shock (can't let go)

Exercise:

Problem:

While taking a bath, a person touches the metal case of a radio. The path through the person to the drainpipe and ground has a resistance of $4000\ \Omega$. What is the smallest voltage on the case of the radio that could cause ventricular fibrillation?

Exercise:

Problem:

Foolishly trying to fish a burning piece of bread from a toaster with a metal butter knife, a man comes into contact with 120-V AC. He does not even feel it since, luckily, he is wearing rubber-soled shoes. What is the minimum resistance of the path the current follows through the person?

Solution:

$1.20 \times 10^5\ \Omega$

Exercise:

Problem:

(a) During surgery, a current as small as $20.0\ \mu\text{A}$ applied directly to the heart may cause ventricular fibrillation. If the resistance of the exposed heart is $300\ \Omega$, what is the smallest voltage that poses this danger? (b) Does your answer imply that special electrical safety precautions are needed?

Exercise:

Problem:

(a) What is the resistance of a 220-V AC short circuit that generates a peak power of 96.8 kW? (b) What would the average power be if the voltage was 120 V AC?

Solution:

(a) $1.00\ \Omega$

(b) 14.4 kW

Exercise:

Problem:

A heart defibrillator passes 10.0 A through a patient's torso for 5.00 ms in an attempt to restore normal beating. (a) How much charge passed? (b) What voltage was applied if 500 J of energy was dissipated? (c) What was the path's resistance? (d) Find the temperature increase caused in the 8.00 kg of affected tissue.

Exercise:

Problem: Integrated Concepts

A short circuit in a 120-V appliance cord has a $0.500\text{-}\Omega$ resistance. Calculate the temperature rise of the 2.00 g of surrounding materials, assuming their specific heat capacity is $0.200 \text{ cal/g} \cdot^{\text{o}} \text{C}$ and that it takes 0.0500 s for a circuit breaker to interrupt the current. Is this likely to be damaging?

Solution:

Temperature increases 860^{o} C. It is very likely to be damaging.

Exercise:

Problem: Construct Your Own Problem

Consider a person working in an environment where electric currents might pass through her body. Construct a problem in which you calculate the resistance of insulation needed to protect the person from harm. Among the things to be considered are the voltage to which the person might be exposed, likely body resistance (dry, wet, …), and acceptable currents (safe but sensed, safe and unfelt, …).

Glossary

thermal hazard
: a hazard in which electric current causes undesired thermal effects

shock hazard
: when electric current passes through a person

short circuit
: also known as a “short,” a low-resistance path between terminals of a voltage source

microshock sensitive
: a condition in which a person’s skin resistance is bypassed, possibly by a medical procedure, rendering the person vulnerable to electrical shock at currents about 1/1000 the normally required level

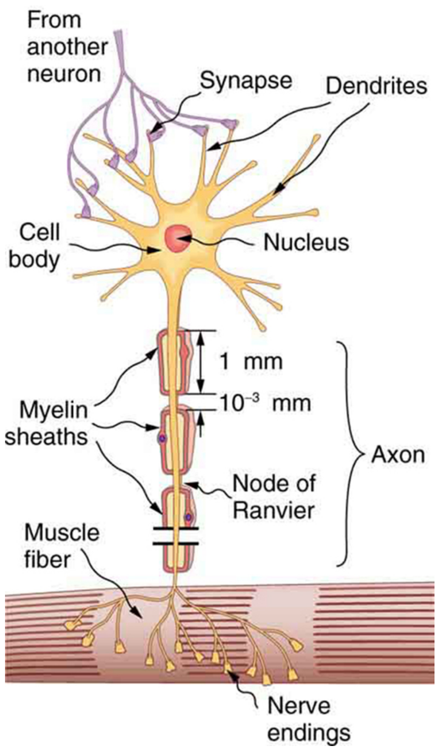
Nerve Conduction–Electrocardiograms

- Explain the process by which electric signals are transmitted along a neuron.
- Explain the effects myelin sheaths have on signal propagation.
- Explain what the features of an ECG signal indicate.

Nerve Conduction

Electric currents in the vastly complex system of billions of nerves in our body allow us to sense the world, control parts of our body, and think. These are representative of the three major functions of nerves. First, nerves carry messages from our sensory organs and others to the central nervous system, consisting of the brain and spinal cord. Second, nerves carry messages from the central nervous system to muscles and other organs. Third, nerves transmit and process signals within the central nervous system. The sheer number of nerve cells and the incredibly greater number of connections between them makes this system the subtle wonder that it is. **Nerve conduction** is a general term for electrical signals carried by nerve cells. It is one aspect of **bioelectricity**, or electrical effects in and created by biological systems.

Nerve cells, properly called *neurons*, look different from other cells—they have tendrils, some of them many centimeters long, connecting them with other cells. (See [link].) Signals arrive at the cell body across *synapses* or through *dendrites*, stimulating the neuron to generate its own signal, sent along its long *axon* to other nerve or muscle cells. Signals may arrive from many other locations and be transmitted to yet others, conditioning the synapses by use, giving the system its complexity and its ability to learn.

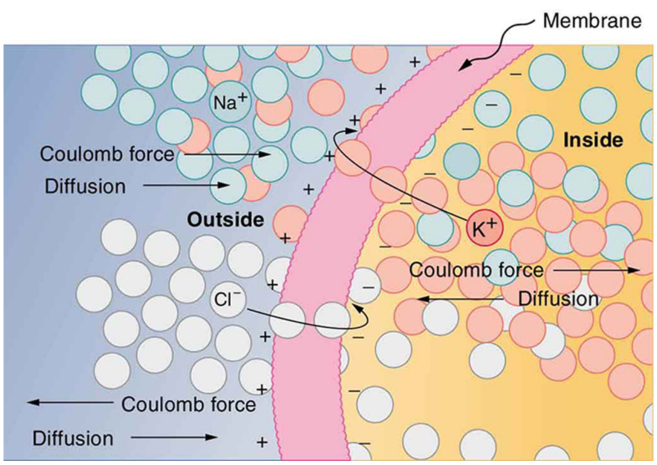


A neuron with its dendrites and long axon. Signals in the form of electric currents reach the cell body through dendrites and across synapses, stimulating the neuron to generate its own signal sent down the axon. The number of interconnections can be far greater than shown here.

The method by which these electric currents are generated and transmitted is more complex than the simple movement of free charges in a conductor,

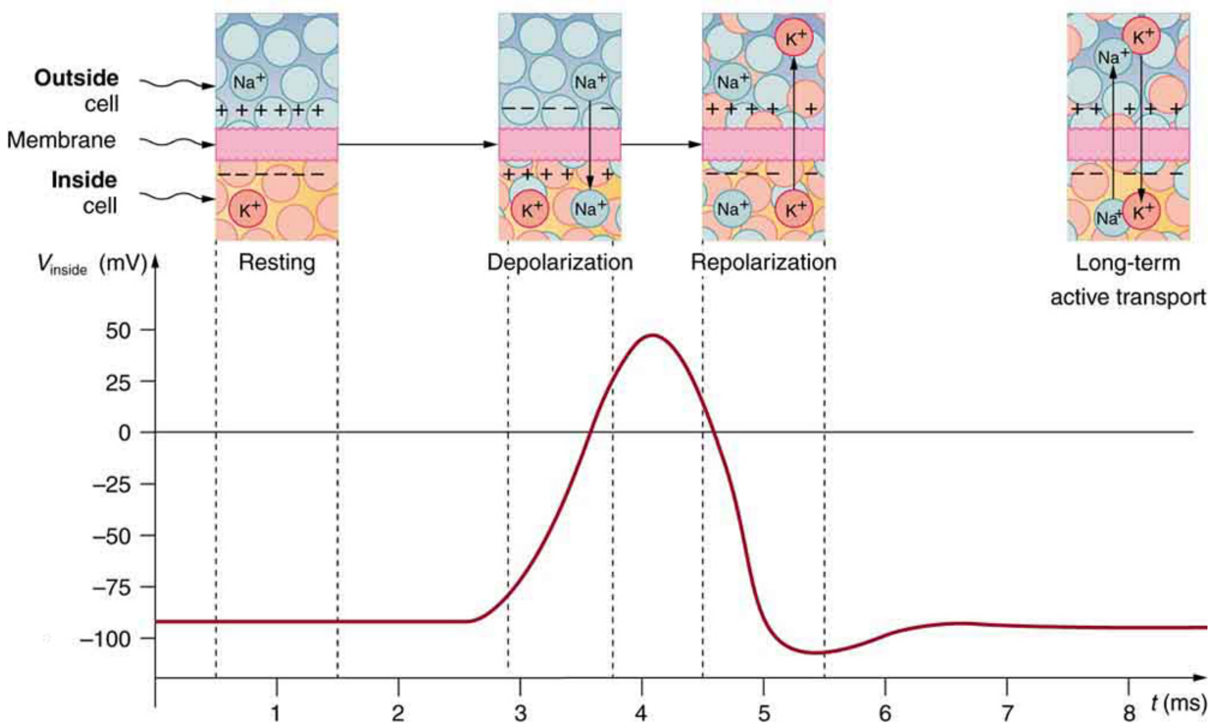
but it can be understood with principles already discussed in this text. The most important of these are the Coulomb force and diffusion.

[link] illustrates how a voltage (potential difference) is created across the cell membrane of a neuron in its resting state. This thin membrane separates electrically neutral fluids having differing concentrations of ions, the most important varieties being Na^+, K^+, and Cl^- (these are sodium, potassium, and chlorine ions with single plus or minus charges as indicated). As discussed in Molecular Transport Phenomena: Diffusion, Osmosis, and Related Processes, free ions will diffuse from a region of high concentration to one of low concentration. But the cell membrane is **semipermeable**, meaning that some ions may cross it while others cannot. In its resting state, the cell membrane is permeable to K^+ and Cl^-, and impermeable to Na^+. Diffusion of K^+ and Cl^- thus creates the layers of positive and negative charge on the outside and inside of the membrane. The Coulomb force prevents the ions from diffusing across in their entirety. Once the charge layer has built up, the repulsion of like charges prevents more from moving across, and the attraction of unlike charges prevents more from leaving either side. The result is two layers of charge right on the membrane, with diffusion being balanced by the Coulomb force. A tiny fraction of the charges move across and the fluids remain neutral (other ions are present), while a separation of charge and a voltage have been created across the membrane.



The semipermeable membrane of a

cell has different concentrations of ions inside and out. Diffusion moves the K^+ and Cl^- ions in the direction shown, until the Coulomb force halts further transfer. This results in a layer of positive charge on the outside, a layer of negative charge on the inside, and thus a voltage across the cell membrane. The membrane is normally impermeable to Na^+.



An action potential is the pulse of voltage inside a nerve cell graphed here. It is caused by movements of ions across the cell membrane as shown. Depolarization occurs when a stimulus makes the membrane permeable to Na^+ ions. Repolarization follows as the membrane

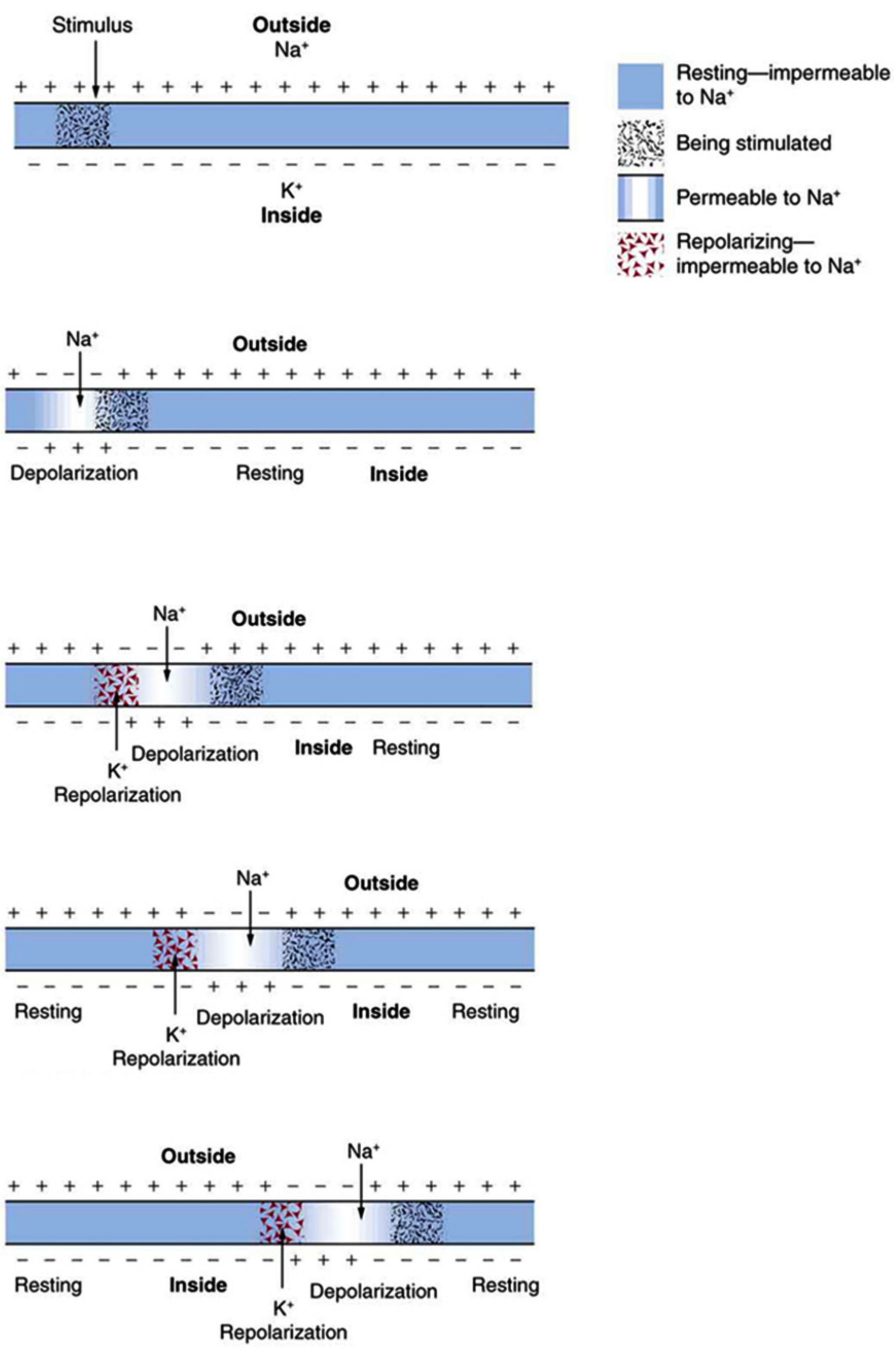
again becomes impermeable to $\mathrm{Na^+}$, and $\mathrm{K^+}$ moves from high to low concentration. In the long term, active transport slowly maintains the concentration differences, but the cell may fire hundreds of times in rapid succession without seriously depleting them.

The separation of charge creates a potential difference of 70 to 90 mV across the cell membrane. While this is a small voltage, the resulting electric field ($E = V/d$) across the only 8-nm-thick membrane is immense (on the order of 11 MV/m!) and has fundamental effects on its structure and permeability. Now, if the exterior of a neuron is taken to be at 0 V, then the interior has a *resting potential* of about –90 mV. Such voltages are created across the membranes of almost all types of animal cells but are largest in nerve and muscle cells. In fact, fully 25% of the energy used by cells goes toward creating and maintaining these potentials.

Electric currents along the cell membrane are created by any stimulus that changes the membrane's permeability. The membrane thus temporarily becomes permeable to $\mathrm{Na^+}$, which then rushes in, driven both by diffusion and the Coulomb force. This inrush of $\mathrm{Na^+}$ first neutralizes the inside membrane, or *depolarizes* it, and then makes it slightly positive. The depolarization causes the membrane to again become impermeable to $\mathrm{Na^+}$, and the movement of $\mathrm{K^+}$ quickly returns the cell to its resting potential, or *repolarizes* it. This sequence of events results in a voltage pulse, called the *action potential*. (See [link].) Only small fractions of the ions move, so that the cell can fire many hundreds of times without depleting the excess concentrations of $\mathrm{Na^+}$ and $\mathrm{K^+}$. Eventually, the cell must replenish these ions to maintain the concentration differences that create bioelectricity. This sodium-potassium pump is an example of *active transport*, wherein cell energy is used to move ions across membranes against diffusion gradients and the Coulomb force.

The action potential is a voltage pulse at one location on a cell membrane. How does it get transmitted along the cell membrane, and in particular down an axon, as a nerve impulse? The answer is that the changing voltage and electric fields affect the permeability of the adjacent cell membrane, so

that the same process takes place there. The adjacent membrane depolarizes, affecting the membrane further down, and so on, as illustrated in [link]. Thus the action potential stimulated at one location triggers a *nerve impulse* that moves slowly (about 1 m/s) along the cell membrane.



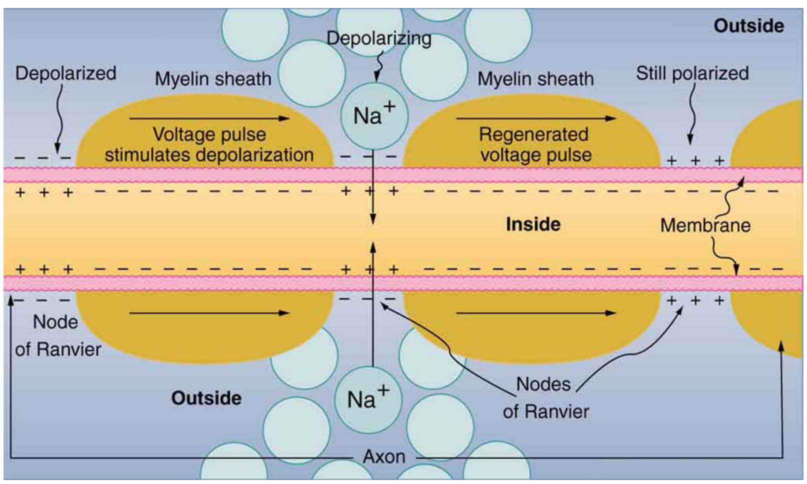
A nerve impulse is the propagation of an action potential along a cell membrane. A stimulus causes an action potential at one location, which changes the permeability of the adjacent membrane, causing an action potential there. This in turn affects the membrane further down, so that the action potential moves slowly (in electrical terms) along the cell membrane. Although the impulse is due to Na^+ and K^+ going across the membrane, it is equivalent to a wave of charge moving along the outside and inside of the membrane.

Some axons, like that in [link], are sheathed with *myelin*, consisting of fat-containing cells. [link] shows an enlarged view of an axon having myelin sheaths characteristically separated by unmyelinated gaps (called nodes of Ranvier). This arrangement gives the axon a number of interesting properties. Since myelin is an insulator, it prevents signals from jumping between adjacent nerves (cross talk). Additionally, the myelinated regions transmit electrical signals at a very high speed, as an ordinary conductor or resistor would. There is no action potential in the myelinated regions, so that no cell energy is used in them. There is an IR signal loss in the myelin, but the signal is regenerated in the gaps, where the voltage pulse triggers the action potential at full voltage. So a myelinated axon transmits a nerve impulse faster, with less energy consumption, and is better protected from cross talk than an unmyelinated one. Not all axons are myelinated, so that cross talk and slow signal transmission are a characteristic of the normal operation of these axons, another variable in the nervous system.

The degeneration or destruction of the myelin sheaths that surround the nerve fibers impairs signal transmission and can lead to numerous neurological effects. One of the most prominent of these diseases comes from the body’s own immune system attacking the myelin in the central nervous system—multiple sclerosis. MS symptoms include fatigue, vision problems, weakness of arms and legs, loss of balance, and tingling or

numbness in one’s extremities (neuropathy). It is more apt to strike younger adults, especially females. Causes might come from infection, environmental or geographic affects, or genetics. At the moment there is no known cure for MS.

Most animal cells can fire or create their own action potential. Muscle cells contract when they fire and are often induced to do so by a nerve impulse. In fact, nerve and muscle cells are physiologically similar, and there are even hybrid cells, such as in the heart, that have characteristics of both nerves and muscles. Some animals, like the infamous electric eel (see [link]), use muscles ganged so that their voltages add in order to create a shock great enough to stun prey.

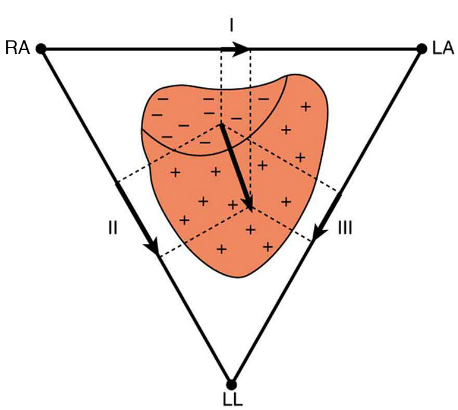


Propagation of a nerve impulse down a myelinated axon, from left to right. The signal travels very fast and without energy input in the myelinated regions, but it loses voltage. It is regenerated in the gaps. The signal moves faster than in unmyelinated axons and is insulated from signals in other nerves, limiting cross talk.

An electric eel flexes its muscles to create a voltage that stuns prey. (credit: chrisbb, Flickr)

Electrocardiograms

Just as nerve impulses are transmitted by depolarization and repolarization of adjacent membrane, the depolarization that causes muscle contraction can also stimulate adjacent muscle cells to depolarize (fire) and contract. Thus, a depolarization wave can be sent across the heart, coordinating its rhythmic contractions and enabling it to perform its vital function of propelling blood through the circulatory system. [link] is a simplified graphic of a depolarization wave spreading across the heart from the *sinoarterial (SA) node***,** the heart’s natural pacemaker.

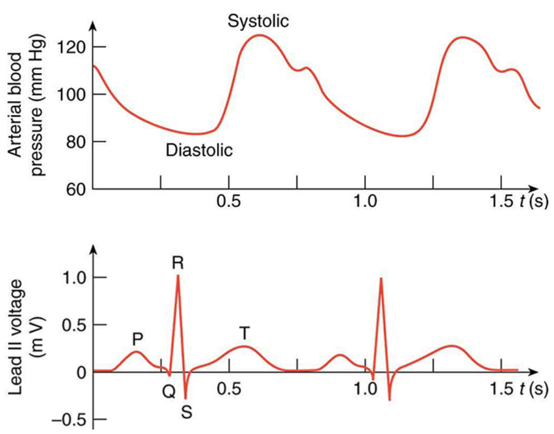


The outer surface of the heart changes from positive to negative during depolarization. This wave of depolarization is spreading from the top of the heart and is represented by a vector pointing in the direction of the wave. This vector is a voltage (potential difference) vector. Three electrodes, labeled RA, LA, and LL, are placed on the patient. Each pair (called leads I, II, and III) measures a component of the depolarization vector and is graphed in an ECG.

An **electrocardiogram (ECG)** is a record of the voltages created by the wave of depolarization and subsequent repolarization in the heart. Voltages between pairs of electrodes placed on the chest are vector components of the voltage wave on the heart. Standard ECGs have 12 or more electrodes, but only three are shown in [link] for clarity. Decades ago, three-electrode ECGs were performed by placing electrodes on the left and right arms and the left leg. The voltage between the right arm and the left leg is called the *lead II potential* and is the most often graphed. We shall examine the lead II potential as an indicator of heart-muscle function and see that it is coordinated with arterial blood pressure as well.

Heart function and its four-chamber action are explored in Viscosity and Laminar Flow; Poiseuille's Law. Basically, the right and left atria receive blood from the body and lungs, respectively, and pump the blood into the ventricles. The right and left ventricles, in turn, pump blood through the lungs and the rest of the body, respectively. Depolarization of the heart muscle causes it to contract. After contraction it is repolarized to ready it for the next beat. The ECG measures components of depolarization and repolarization of the heart muscle and can yield significant information on the functioning and malfunctioning of the heart.

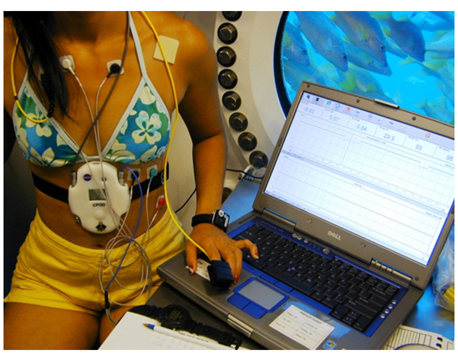
[link] shows an ECG of the lead II potential and a graph of the corresponding arterial blood pressure. The major features are labeled P, Q, R, S, and T. The *P wave* is generated by the depolarization and contraction of the atria as they pump blood into the ventricles. The *QRS complex* is created by the depolarization of the ventricles as they pump blood to the lungs and body. Since the shape of the heart and the path of the depolarization wave are not simple, the QRS complex has this typical shape and time span. The lead II QRS signal also masks the repolarization of the atria, which occur at the same time. Finally, the *T wave* is generated by the repolarization of the ventricles and is followed by the next P wave in the next heartbeat. Arterial blood pressure varies with each part of the heartbeat, with systolic (maximum) pressure occurring closely after the QRS complex, which signals contraction of the ventricles.



A lead II ECG with

corresponding arterial blood pressure. The QRS complex is created by the depolarization and contraction of the ventricles and is followed shortly by the maximum or systolic blood pressure. See text for further description.

Taken together, the 12 leads of a state-of-the-art ECG can yield a wealth of information about the heart. For example, regions of damaged heart tissue, called infarcts, reflect electrical waves and are apparent in one or more lead potentials. Subtle changes due to slight or gradual damage to the heart are most readily detected by comparing a recent ECG to an older one. This is particularly the case since individual heart shape, size, and orientation can cause variations in ECGs from one individual to another. ECG technology has advanced to the point where a portable ECG monitor with a liquid crystal instant display and a printer can be carried to patients' homes or used in emergency vehicles. See [link].



This NASA scientist and NEEMO 5 aquanaut’s heart rate and other vital signs

are being recorded by a portable device while living in an underwater habitat. (credit: NASA, Life Sciences Data Archive at Johnson Space Center, Houston, Texas)

Note:
PhET Explorations: Neuron
Stimulate a neuron and monitor what happens. Pause, rewind, and move forward in time in order to observe the ions as they move across the neuron membrane.
https://phet.colorado.edu/sims/html/neuron/latest/neuron_en.html

Section Summary

- Electric potentials in neurons and other cells are created by ionic concentration differences across semipermeable membranes.
- Stimuli change the permeability and create action potentials that propagate along neurons.
- Myelin sheaths speed this process and reduce the needed energy input.
- This process in the heart can be measured with an electrocardiogram (ECG).

Conceptual Questions

Exercise:

Problem:

Note that in [link], both the concentration gradient and the Coulomb force tend to move Na^+ ions into the cell. What prevents this?

Exercise:

Problem:

Define depolarization, repolarization, and the action potential.

Exercise:

Problem:

Explain the properties of myelinated nerves in terms of the insulating properties of myelin.

Problems & Exercises

Exercise:

Problem: Integrated Concepts

Use the ECG in [link] to determine the heart rate in beats per minute assuming a constant time between beats.

Solution:

80 beats/minute

Exercise:

Problem: Integrated Concepts

(a) Referring to [link], find the time systolic pressure lags behind the middle of the QRS complex. (b) Discuss the reasons for the time lag.

Glossary

nerve conduction
: the transport of electrical signals by nerve cells

bioelectricity
: electrical effects in and created by biological systems

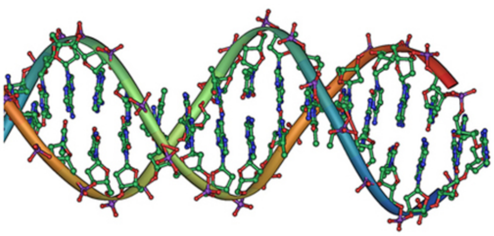
semipermeable
: property of a membrane that allows only certain types of ions to cross it

electrocardiogram (ECG)
: usually abbreviated ECG, a record of voltages created by depolarization and repolarization, especially in the heart

Electric Forces in Biology

- Describe how a water molecule is polar.
- Explain electrostatic screening by a water molecule within a living cell.

Classical electrostatics has an important role to play in modern molecular biology. Large molecules such as proteins, nucleic acids, and so on—so important to life—are usually electrically charged. DNA itself is highly charged; it is the electrostatic force that not only holds the molecule together but gives the molecule structure and strength. [link] is a schematic of the DNA double helix.



DNA is a highly charged molecule. The DNA double helix shows the two coiled strands each containing a row of nitrogenous bases, which "code" the genetic information needed by a living organism. The strands are connected by bonds between pairs of bases. While pairing combinations between certain bases are fixed (C-G and A-T), the sequence of nucleotides in the strand varies. (credit: Jerome Walker)

The four nucleotide bases are given the symbols A (adenine), C (cytosine), G (guanine), and T (thymine). The order of the four bases varies in each strand, but the pairing between bases is always the same. C and G are always paired and A and T are always paired, which helps to preserve the order of bases in cell division (mitosis) so as to pass on the correct genetic information. Since the Coulomb force drops with distance ($F \propto 1/r^2$), the distances between the base pairs must be small enough that the electrostatic force is sufficient to hold them together.

DNA is a highly charged molecule, with about $2q_{\mathrm{e}}$ (fundamental charge) per 0.3×10^{-9} m. The distance separating the two strands that make up the DNA structure is about 1 nm, while the distance separating the individual atoms within each base is about 0.3 nm.

One might wonder why electrostatic forces do not play a larger role in biology than they do if we have so many charged molecules. The reason is that the electrostatic force is “diluted” due to **screening** between molecules. This is due to the presence of other charges in the cell.

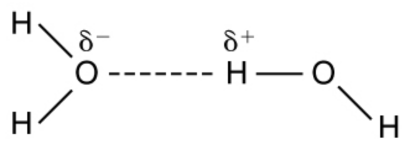
Polarity of Water Molecules

The best example of this charge screening is the water molecule, represented as H_2O. Water is a strongly **polar molecule**. Its 10 electrons (8 from the oxygen atom and 2 from the two hydrogen atoms) tend to remain closer to the oxygen nucleus than the hydrogen nuclei. This creates two centers of equal and opposite charges—what is called a **dipole**, as illustrated in [link]. The magnitude of the dipole is called the dipole moment.

These two centers of charge will terminate some of the electric field lines coming from a free charge, as on a DNA molecule. This results in a reduction in the strength of the **Coulomb interaction**. One might say that screening makes the Coulomb force a short range force rather than long range.

Other ions of importance in biology that can reduce or screen Coulomb interactions are Na^+, and K^+, and Cl^-. These ions are located both inside and outside of living cells. The movement of these ions through cell membranes is crucial to the motion of nerve impulses through nerve axons.

Recent studies of electrostatics in biology seem to show that electric fields in cells can be extended over larger distances, in spite of screening, by “microtubules” within the cell. These microtubules are hollow tubes composed of proteins that guide the movement of chromosomes when cells divide, the motion of other organisms within the cell, and provide mechanisms for motion of some cells (as motors).



This schematic shows water (H_2O) as a polar molecule. Unequal sharing of electrons between the oxygen (O) and hydrogen (H) atoms leads to a net separation of positive and negative charge—forming a dipole. The symbols δ^- and δ^+ indicate that the oxygen side of the H_2O molecule tends to be more negative, while the hydrogen ends tend

to be more positive.
This leads to an
attraction of
opposite charges
between molecules.

Section Summary

- Many molecules in living organisms, such as DNA, carry a charge.
- An uneven distribution of the positive and negative charges within a polar molecule produces a dipole.
- The effect of a Coulomb field generated by a charged object may be reduced or blocked by other nearby charged objects.
- Biological systems contain water, and because water molecules are polar, they have a strong effect on other molecules in living systems.

Conceptual Question

Exercise:

Problem:

A cell membrane is a thin layer enveloping a cell. The thickness of the membrane is much less than the size of the cell. In a static situation the membrane has a charge distribution of -2.5×10^{-6}C/m ^{2} on its inner surface and $+2.5 \times 10^{-6}$ C/m^2 on its outer surface. Draw a diagram of the cell and the surrounding cell membrane. Include on this diagram the charge distribution and the corresponding electric field. Is there any electric field inside the cell? Is there any electric field outside the cell?

Glossary

dipole

a molecule’s lack of symmetrical charge distribution, causing one side to be more positive and another to be more negative

polar molecule
: a molecule with an asymmetrical distribution of positive and negative charge

screening
: the dilution or blocking of an electrostatic force on a charged object by the presence of other charges nearby

Coulomb interaction
: the interaction between two charged particles generated by the Coulomb forces they exert on one another

Introduction to the Physics of Hearing
class="introduction"

This tree fell some time ago. When it fell, atoms in the air were disturbed. Physicists would call this disturbance sound whether someone was around to hear it or not. (credit: B.A. Bowen Photography)

If a tree falls in the forest and no one is there to hear it, does it make a sound? The answer to this old philosophical question depends on how you define sound. If sound only exists when someone is around to perceive it, then there was no sound. However, if we define sound in terms of physics; that is, a disturbance of the atoms in matter transmitted from its origin outward (in other words, a wave), then there *was* a sound, even if nobody was around to hear it.

Such a wave is the physical phenomenon we call *sound.* Its perception is hearing. Both the physical phenomenon and its perception are interesting and will be considered in this text. We shall explore both sound and hearing; they are related, but are not the same thing. We will also explore the many practical uses of sound waves, such as in medical imaging.

Sound

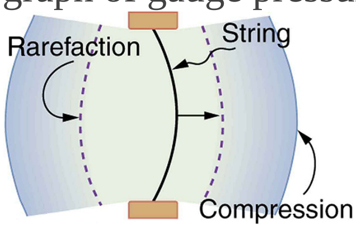
- Define sound and hearing.
- Describe sound as a longitudinal wave.

This glass has been shattered by a high-intensity sound wave of the same frequency as the resonant frequency of the glass. While the sound is not visible, the effects of the sound prove its existence. (credit: ||read||, Flickr)

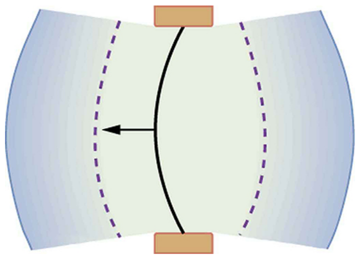
Sound can be used as a familiar illustration of waves. Because hearing is one of our most important senses, it is interesting to see how the physical properties of sound correspond to our perceptions of it. **Hearing** is the perception of sound, just as vision is the perception of visible light. But sound has important applications beyond hearing. Ultrasound, for example, is not heard but can be employed to form medical images and is also used in treatment.

The physical phenomenon of **sound** is defined to be a disturbance of matter that is transmitted from its source outward. Sound is a wave. On the atomic scale, it is a disturbance of atoms that is far more ordered than their thermal motions. In many instances, sound is a periodic wave, and the atoms undergo simple harmonic motion. In this text, we shall explore such periodic sound waves.

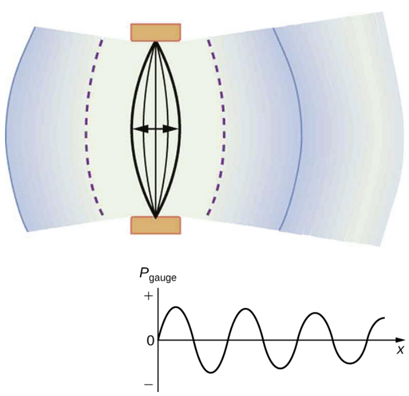
A vibrating string produces a sound wave as illustrated in [link], [link], and [link]. As the string oscillates back and forth, it transfers energy to the air, mostly as thermal energy created by turbulence. But a small part of the string's energy goes into compressing and expanding the surrounding air, creating slightly higher and lower local pressures. These compressions (high pressure regions) and rarefactions (low pressure regions) move out as longitudinal pressure waves having the same frequency as the string—they are the disturbance that is a sound wave. (Sound waves in air and most fluids are longitudinal, because fluids have almost no shear strength. In solids, sound waves can be both transverse and longitudinal.) [link] shows a graph of gauge pressure versus distance from the vibrating string.



A vibrating string moving to the right compresses the air in front of it and expands the air behind it.



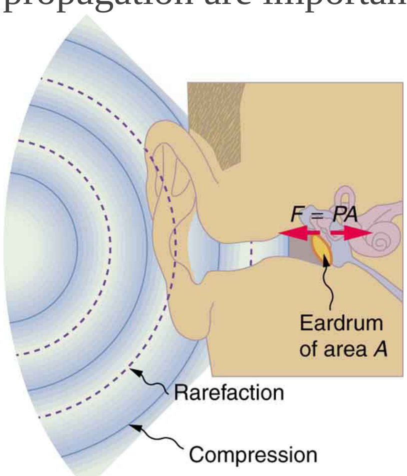
As the string moves to the left, it creates another compression and rarefaction as the ones on the right move away from the string.



After many vibrations, there are a series of compressions and rarefactions moving out from the string as a sound wave. The graph shows gauge pressure versus

distance from the source. Pressures vary only slightly from atmospheric for ordinary sounds.

The amplitude of a sound wave decreases with distance from its source, because the energy of the wave is spread over a larger and larger area. But it is also absorbed by objects, such as the eardrum in [link], and converted to thermal energy by the viscosity of air. In addition, during each compression a little heat transfers to the air and during each rarefaction even less heat transfers from the air, so that the heat transfer reduces the organized disturbance into random thermal motions. (These processes can be viewed as a manifestation of the second law of thermodynamics presented in Introduction to the Second Law of Thermodynamics: Heat Engines and Their Efficiency.) Whether the heat transfer from compression to rarefaction is significant depends on how far apart they are—that is, it depends on wavelength. Wavelength, frequency, amplitude, and speed of propagation are important for sound, as they are for all waves.



Sound wave compressions and rarefactions travel up the ear canal and

force the eardrum to vibrate. There is a net force on the eardrum, since the sound wave pressures differ from the atmospheric pressure found behind the eardrum. A complicated mechanism converts the vibrations to nerve impulses, which are perceived by the person.

Note:
PhET Explorations: Wave Interference
WMake waves with a dripping faucet, audio speaker, or laser! Add a second source or a pair of slits to create an interference pattern.
https://archive.cnx.org/specials/2fe7ad15-b00e-4402-b068-ff503985a18f/wave-interference/

Section Summary

- Sound is a disturbance of matter that is transmitted from its source outward.
- Sound is one type of wave.

- Hearing is the perception of sound.

Glossary

sound
: a disturbance of matter that is transmitted from its source outward

hearing
: the perception of sound

Speed of Sound, Frequency, and Wavelength

- Define pitch.
- Describe the relationship between the speed of sound, its frequency, and its wavelength.
- Describe the effects on the speed of sound as it travels through various media.
- Describe the effects of temperature on the speed of sound.

When a firework explodes, the light energy is perceived before the sound energy. Sound travels more slowly than light does. (credit: Dominic Alves, Flickr)

Sound, like all waves, travels at a certain speed and has the properties of frequency and wavelength. You can observe direct evidence of the speed of sound while watching a fireworks display. The flash of an explosion is seen well before its sound is heard, implying both that sound travels at a finite speed and that it is much slower than light. You can also directly sense the frequency of a sound. Perception of frequency is called **pitch**. The wavelength of sound is not directly sensed, but indirect evidence is found in the correlation of the size of musical instruments with their pitch. Small

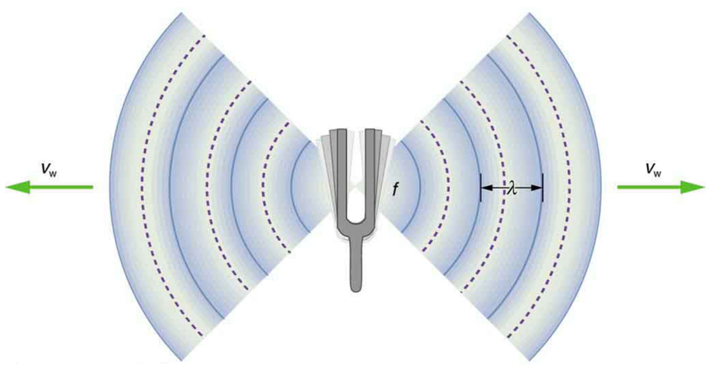
instruments, such as a piccolo, typically make high-pitch sounds, while large instruments, such as a tuba, typically make low-pitch sounds. High pitch means small wavelength, and the size of a musical instrument is directly related to the wavelengths of sound it produces. So a small instrument creates short-wavelength sounds. Similar arguments hold that a large instrument creates long-wavelength sounds.

The relationship of the speed of sound, its frequency, and wavelength is the same as for all waves:

Equation:

$$v_{\mathrm{w}} = f\lambda,$$

where v_{w} is the speed of sound, f is its frequency, and λ is its wavelength. The wavelength of a sound is the distance between adjacent identical parts of a wave—for example, between adjacent compressions as illustrated in [link]. The frequency is the same as that of the source and is the number of waves that pass a point per unit time.



A sound wave emanates from a source vibrating at a frequency f, propagates at v_{w}, and has a wavelength λ.

[link] makes it apparent that the speed of sound varies greatly in different media. The speed of sound in a medium is determined by a combination of the medium’s rigidity (or compressibility in gases) and its density. The

more rigid (or less compressible) the medium, the faster the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is directly proportional to the stiffness of the oscillating object. The greater the density of a medium, the slower the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is inversely proportional to the mass of the oscillating object. The speed of sound in air is low, because air is compressible. Because liquids and solids are relatively rigid and very difficult to compress, the speed of sound in such media is generally greater than in gases.

Medium	**v_w(m/s)**
Gases at $0^{\circ}C$	
Air	331
Carbon dioxide	259
Oxygen	316
Helium	965
Hydrogen	1290
Liquids at $20^{\circ}C$	
Ethanol	1160
Mercury	1450
Water, fresh	1480

Medium	**v_w(m/s)**
Sea water	1540
Human tissue	1540
Solids (longitudinal or bulk)	
Vulcanized rubber	54
Polyethylene	920
Marble	3810
Glass, Pyrex	5640
Lead	1960
Aluminum	5120
Steel	5960

Speed of Sound in Various Media

Earthquakes, essentially sound waves in Earth's crust, are an interesting example of how the speed of sound depends on the rigidity of the medium. Earthquakes have both longitudinal and transverse components, and these travel at different speeds. The bulk modulus of granite is greater than its shear modulus. For that reason, the speed of longitudinal or pressure waves (P-waves) in earthquakes in granite is significantly higher than the speed of transverse or shear waves (S-waves). Both components of earthquakes travel slower in less rigid material, such as sediments. P-waves have speeds of 4 to 7 km/s, and S-waves correspondingly range in speed from 2 to 5 km/s, both being faster in more rigid material. The P-wave gets progressively farther ahead of the S-wave as they travel through Earth's crust. The time between the P- and S-waves is routinely used to determine the distance to their source, the epicenter of the earthquake.

The speed of sound is affected by temperature in a given medium. For air at sea level, the speed of sound is given by

Equation:

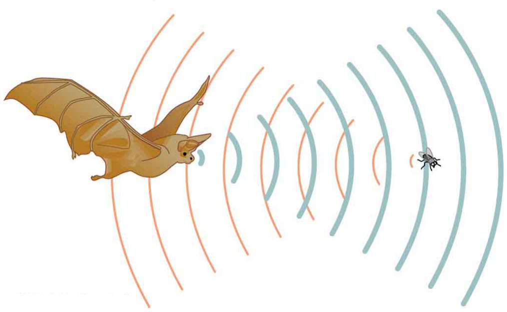
$$v_{\mathrm{w}} = (331\ \mathrm{m/s})\sqrt{\frac{T}{273\ \mathrm{K}}},$$

where the temperature (denoted as T) is in units of kelvin. The speed of sound in gases is related to the average speed of particles in the gas, v_{rms}, and that

Equation:

$$v_{\mathrm{rms}} = \sqrt{\frac{3\,kT}{m}},$$

where k is the Boltzmann constant ($1.38 \times 10^{-23}\ \mathrm{J/K}$) and m is the mass of each (identical) particle in the gas. So, it is reasonable that the speed of sound in air and other gases should depend on the square root of temperature. While not negligible, this is not a strong dependence. At 0ºC, the speed of sound is 331 m/s, whereas at 20.0ºC it is 343 m/s, less than a 4% increase. [link] shows a use of the speed of sound by a bat to sense distances. Echoes are also used in medical imaging.



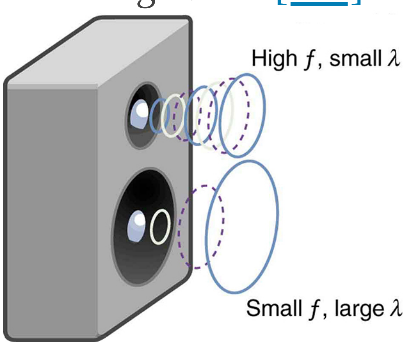
A bat uses sound echoes to find its way about and to catch prey. The time for the echo to return is directly proportional to the distance.

One of the more important properties of sound is that its speed is nearly independent of frequency. This independence is certainly true in open air for sounds in the audible range of 20 to 20,000 Hz. If this independence were not true, you would certainly notice it for music played by a marching band in a football stadium, for example. Suppose that high-frequency sounds traveled faster—then the farther you were from the band, the more the sound from the low-pitch instruments would lag that from the high-pitch ones. But the music from all instruments arrives in cadence independent of distance, and so all frequencies must travel at nearly the same speed. Recall that

Equation:

$$v_{\text{w}} = f\lambda.$$

In a given medium under fixed conditions, v_{w} is constant, so that there is a relationship between f and λ; the higher the frequency, the smaller the wavelength. See [link] and consider the following example.



Because they travel at the same speed in a given medium, low-frequency sounds must have a greater wavelength than high-frequency sounds.

Here, the lower-frequency sounds are emitted by the large speaker, called a woofer, while the higher-frequency sounds are emitted by the small speaker, called a tweeter.

Example:
Calculating Wavelengths: What Are the Wavelengths of Audible Sounds?

Calculate the wavelengths of sounds at the extremes of the audible range, 20 and 20,000 Hz, in 30.0ºC air. (Assume that the frequency values are accurate to two significant figures.)

Strategy

To find wavelength from frequency, we can use $v_{\text{w}} = f\lambda$.

Solution

1. Identify knowns. The value for v_{w}, is given by
   Equation:
   $$v_{\text{w}} = (331\ \text{m/s})\sqrt{\frac{T}{273\ \text{K}}}.$$
2. Convert the temperature into kelvin and then enter the temperature into the equation
   Equation:
   $$v_{\text{w}} = (331\ \text{m/s})\sqrt{\frac{303\ \text{K}}{273\ \text{K}}} = 348.7\ \text{m/s}.$$

3. Solve the relationship between speed and wavelength for λ:
   Equation:

$$\lambda = \frac{v_{\text{w}}}{f}.$$

4. Enter the speed and the minimum frequency to give the maximum wavelength:
   Equation:

$$\lambda_{\text{max}} = \frac{348.7\ \text{m/s}}{20\ \text{Hz}} = 17\ \text{m}.$$

5. Enter the speed and the maximum frequency to give the minimum wavelength:
   Equation:

$$\lambda_{\text{min}} = \frac{348.7\ \text{m/s}}{20{,}000\ \text{Hz}} = 0.017\ \text{m} = 1.7\ \text{cm}.$$

Discussion
Because the product of f multiplied by λ equals a constant, the smaller f is, the larger λ must be, and vice versa.

The speed of sound can change when sound travels from one medium to another. However, the frequency usually remains the same because it is like a driven oscillation and has the frequency of the original source. If v_{w} changes and f remains the same, then the wavelength λ must change. That is, because $v_{\text{w}} = f\lambda$, the higher the speed of a sound, the greater its wavelength for a given frequency.

Note:
Making Connections: Take-Home Investigation—Voice as a Sound Wave

Suspend a sheet of paper so that the top edge of the paper is fixed and the bottom edge is free to move. You could tape the top edge of the paper to the edge of a table. Gently blow near the edge of the bottom of the sheet and note how the sheet moves. Speak softly and then louder such that the sounds hit the edge of the bottom of the paper, and note how the sheet moves. Explain the effects.

Exercise:
Check Your Understanding

Problem:

Imagine you observe two fireworks explode. You hear the explosion of one as soon as you see it. However, you see the other firework for several milliseconds before you hear the explosion. Explain why this is so.

Solution:

Sound and light both travel at definite speeds. The speed of sound is slower than the speed of light. The first firework is probably very close by, so the speed difference is not noticeable. The second firework is farther away, so the light arrives at your eyes noticeably sooner than the sound wave arrives at your ears.

Exercise:
Check Your Understanding

Problem:

You observe two musical instruments that you cannot identify. One plays high-pitch sounds and the other plays low-pitch sounds. How could you determine which is which without hearing either of them play?

Solution:

Compare their sizes. High-pitch instruments are generally smaller than low-pitch instruments because they generate a smaller wavelength.

Section Summary

The relationship of the speed of sound v_{w}, its frequency f, and its wavelength λ is given by

Equation:

$$v_{\text{w}} = f\lambda,$$

which is the same relationship given for all waves.

In air, the speed of sound is related to air temperature T by

Equation:

$$v_{\text{w}} = (331\text{ m/s})\sqrt{\frac{T}{273\text{ K}}}.$$

v_{w} is the same for all frequencies and wavelengths.

Conceptual Questions

Exercise:

Problem:

How do sound vibrations of atoms differ from thermal motion?

Exercise:

Problem:

When sound passes from one medium to another where its propagation speed is different, does its frequency or wavelength change? Explain your answer briefly.

Problems & Exercises

Exercise:

Problem:

When poked by a spear, an operatic soprano lets out a 1200-Hz shriek. What is its wavelength if the speed of sound is 345 m/s?

Solution:

0.288 m

Exercise:

Problem:

What frequency sound has a 0.10-m wavelength when the speed of sound is 340 m/s?

Exercise:

Problem:

Calculate the speed of sound on a day when a 1500 Hz frequency has a wavelength of 0.221 m.

Solution:

332 m/s

Exercise:

Problem:

(a) What is the speed of sound in a medium where a 100-kHz frequency produces a 5.96-cm wavelength? (b) Which substance in [link] is this likely to be?

Exercise:

Problem:

Show that the speed of sound in 20.0°C air is 343 m/s, as claimed in the text.

Solution:
Equation:

$$\begin{aligned} v_{\text{w}} &= (331\text{ m/s})\sqrt{\frac{T}{273\text{ K}}} = (331\text{ m/s})\sqrt{\frac{293\text{ K}}{273\text{ K}}} \\ &= 343\text{ m/s} \end{aligned}$$

Exercise:

Problem:

Air temperature in the Sahara Desert can reach 56.0°C (about 134°F). What is the speed of sound in air at that temperature?

Exercise:

Problem:

Dolphins make sounds in air and water. What is the ratio of the wavelength of a sound in air to its wavelength in seawater? Assume air temperature is 20.0°C.

Solution:

0.223

Exercise:

Problem:

A sonar echo returns to a submarine 1.20 s after being emitted. What is the distance to the object creating the echo? (Assume that the submarine is in the ocean, not in fresh water.)

Exercise:

Problem:

(a) If a submarine’s sonar can measure echo times with a precision of 0.0100 s, what is the smallest difference in distances it can detect? (Assume that the submarine is in the ocean, not in fresh water.)

(b) Discuss the limits this time resolution imposes on the ability of the sonar system to detect the size and shape of the object creating the echo.

Solution:

(a) 7.70 m

(b) This means that sonar is good for spotting and locating large objects, but it isn’t able to resolve smaller objects, or detect the detailed shapes of objects. Objects like ships or large pieces of airplanes can be found by sonar, while smaller pieces must be found by other means.

Exercise:

Problem:

A physicist at a fireworks display times the lag between seeing an explosion and hearing its sound, and finds it to be 0.400 s. (a) How far away is the explosion if air temperature is 24.0°C and if you neglect the time taken for light to reach the physicist? (b) Calculate the distance to the explosion taking the speed of light into account. Note that this distance is negligibly greater.

Exercise:

Problem:

Suppose a bat uses sound echoes to locate its insect prey, 3.00 m away. (See [link].) (a) Calculate the echo times for temperatures of 5.00°C and 35.0°C. (b) What percent uncertainty does this cause for the bat in locating the insect? (c) Discuss the significance of this uncertainty and whether it could cause difficulties for the bat. (In practice, the bat continues to use sound as it closes in, eliminating most of any difficulties imposed by this and other effects, such as motion of the prey.)

Solution:

(a) 18.0 ms, 17.1 ms

(b) 5.00%

(c) This uncertainty could definitely cause difficulties for the bat, if it didn't continue to use sound as it closed in on its prey. A 5% uncertainty could be the difference between catching the prey around the neck or around the chest, which means that it could miss grabbing its prey.

Glossary

pitch
: the perception of the frequency of a sound

Sound Intensity and Sound Level

- Define intensity, sound intensity, and sound pressure level.
- Calculate sound intensity levels in decibels (dB).

Noise on crowded roadways like this one in Delhi makes it hard to hear others unless they shout. (credit: Lingaraj G J, Flickr)

In a quiet forest, you can sometimes hear a single leaf fall to the ground. After settling into bed, you may hear your blood pulsing through your ears. But when a passing motorist has his stereo turned up, you cannot even hear what the person next to you in your car is saying. We are all very familiar with the loudness of sounds and aware that they are related to how energetically the source is vibrating. In cartoons depicting a screaming person (or an animal making a loud noise), the cartoonist often shows an open mouth with a vibrating uvula, the hanging tissue at the back of the mouth, to suggest a loud sound coming from the throat [link]. High noise exposure is hazardous to hearing, and it is common for musicians to have hearing losses that are sufficiently severe that they interfere with the musicians’ abilities to perform. The relevant physical quantity is sound intensity, a concept that is valid for all sounds whether or not they are in the audible range.

Intensity is defined to be the power per unit area carried by a wave. Power is the rate at which energy is transferred by the wave. In equation form, **intensity** I is

Equation:

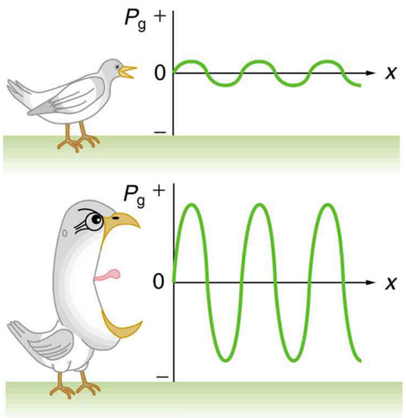
$$I = \frac{P}{A},$$

where P is the power through an area A. The SI unit for I is $\mathrm{W/m^2}$. The intensity of a sound wave is related to its amplitude squared by the following relationship:

Equation:

$$I = \frac{(\Delta p)^2}{2\rho v_{\mathrm{w}}}.$$

Here Δp is the pressure variation or pressure amplitude (half the difference between the maximum and minimum pressure in the sound wave) in units of pascals (Pa) or $\mathrm{N/m^2}$. (We are using a lower case p for pressure to distinguish it from power, denoted by P above.) The energy (as kinetic energy $\frac{mv^2}{2}$) of an oscillating element of air due to a traveling sound wave is proportional to its amplitude squared. In this equation, ρ is the density of the material in which the sound wave travels, in units of $\mathrm{kg/m^3}$, and v_{w} is the speed of sound in the medium, in units of m/s. The pressure variation is proportional to the amplitude of the oscillation, and so I varies as $(\Delta p)^2$ ([link]). This relationship is consistent with the fact that the sound wave is produced by some vibration; the greater its pressure amplitude, the more the air is compressed in the sound it creates.



Graphs of the gauge pressures in two sound waves of different intensities. The more intense sound is produced by a source that has larger-amplitude oscillations and has greater pressure maxima and minima. Because pressures are higher in the greater-intensity sound, it can exert larger forces on the objects it encounters.

Sound intensity levels are quoted in decibels (dB) much more often than sound intensities in watts per meter squared. Decibels are the unit of choice in the scientific literature as well as in the popular media. The reasons for this choice of units are related to how we perceive sounds. How our ears perceive sound can be more accurately described by the logarithm of the

intensity rather than directly to the intensity. The **sound intensity level** β in decibels of a sound having an intensity I in watts per meter squared is defined to be

Equation:

$$\beta\,(\text{dB}) = 10\,\log_{10}\left(\frac{I}{I_0}\right),$$

where $I_0 = 10^{-12}\ \text{W/m}^2$ is a reference intensity. In particular, I_0 is the lowest or threshold intensity of sound a person with normal hearing can perceive at a frequency of 1000 Hz. Sound intensity level is not the same as intensity. Because β is defined in terms of a ratio, it is a unitless quantity telling you the *level* of the sound relative to a fixed standard ($10^{-12}\ \text{W/m}^2$, in this case). The units of decibels (dB) are used to indicate this ratio is multiplied by 10 in its definition. The bel, upon which the decibel is based, is named for Alexander Graham Bell, the inventor of the telephone.

Sound intensity level β (dB)	**Intensity $I(\text{W/m}^2)$**	**Example/effect**
0	1×10^{-12}	Threshold of hearing at 1000 Hz
10	1×10^{-11}	Rustle of leaves
20	1×10^{-10}	Whisper at 1 m distance
30	1×10^{-9}	Quiet home

Sound intensity level β (dB)	**Intensity $I(\text{W/m}^2)$**	**Example/effect**
40	1×10^{-8}	Average home
50	1×10^{-7}	Average office, soft music
60	1×10^{-6}	Normal conversation
70	1×10^{-5}	Noisy office, busy traffic
80	1×10^{-4}	Loud radio, classroom lecture
90	1×10^{-3}	Inside a heavy truck; damage from prolonged exposure[footnote] Several government agencies and health-related professional associations recommend that 85 dB not be exceeded for 8-hour daily exposures in the absence of hearing protection.
100	1×10^{-2}	Noisy factory, siren at 30 m; damage from 8 h per day exposure
110	1×10^{-1}	Damage from 30 min per day exposure
120	1	Loud rock concert, pneumatic chipper at 2 m; threshold of pain
140	1×10^{2}	Jet airplane at 30 m; severe pain, damage in seconds
160	1×10^{4}	Bursting of eardrums

Sound Intensity Levels and Intensities

Sound Intensity Levels and Intensities

The decibel level of a sound having the threshold intensity of $10^{-12}\ \text{W/m}^2$ is $\beta = 0\ \text{dB}$, because $\log_{10} 1 = 0$. That is, the threshold of hearing is 0 decibels. [link] gives levels in decibels and intensities in watts per meter squared for some familiar sounds.

One of the more striking things about the intensities in [link] is that the intensity in watts per meter squared is quite small for most sounds. The ear is sensitive to as little as a trillionth of a watt per meter squared—even more impressive when you realize that the area of the eardrum is only about $1\ \text{cm}^2$, so that only 10^{-16} W falls on it at the threshold of hearing! Air molecules in a sound wave of this intensity vibrate over a distance of less than one molecular diameter, and the gauge pressures involved are less than 10^{-9} atm.

Another impressive feature of the sounds in [link] is their numerical range. Sound intensity varies by a factor of 10^{12} from threshold to a sound that causes damage in seconds. You are unaware of this tremendous range in sound intensity because how your ears respond can be described approximately as the logarithm of intensity. Thus, sound intensity levels in decibels fit your experience better than intensities in watts per meter squared. The decibel scale is also easier to relate to because most people are more accustomed to dealing with numbers such as 0, 53, or 120 than numbers such as 1.00×10^{-11}.

One more observation readily verified by examining [link] or using $I = \frac{(\Delta p)^2}{2\rho v_{\text{w}}}$ is that each factor of 10 in intensity corresponds to 10 dB. For example, a 90 dB sound compared with a 60 dB sound is 30 dB greater, or three factors of 10 (that is, 10^3 times) as intense. Another example is that if one sound is 10^7 as intense as another, it is 70 dB higher. See [link].

I_2/I_1	$\beta_2 - \beta_1$
2.0	3.0 dB
5.0	7.0 dB
10.0	10.0 dB

Ratios of Intensities and Corresponding Differences in Sound Intensity Levels

Example:
Calculating Sound Intensity Levels: Sound Waves
Calculate the sound intensity level in decibels for a sound wave traveling in air at 0ºC and having a pressure amplitude of 0.656 Pa.
Strategy
We are given Δp, so we can calculate I using the equation $I = (\Delta p)^2/(2pv_{\text{w}})^2$. Using I, we can calculate β straight from its definition in $\beta\ (\text{dB}) = 10\ \log_{10}(I/I_0)$.
Solution
(1) Identify knowns:
Sound travels at 331 m/s in air at 0ºC.
Air has a density of $1.29\ \text{kg/m}^3$ at atmospheric pressure and 0ºC.
(2) Enter these values and the pressure amplitude into $I = (\Delta p)^2/\ (2\rho v_{\text{w}})$:
Equation:

$$I = \frac{(\Delta p)^2}{2\rho v_{\text{w}}} = \frac{(0.656\ \text{Pa})^2}{2\left(1.29\ \text{kg/m}^3\right)(331\ \text{m/s})} = 5.04 \times 10^{-4}\ \text{W/m}^2.$$

(3) Enter the value for I and the known value for I_0 into $\beta\ (\text{dB}) = 10\ \log_{10}(I/I_0)$. Calculate to find the sound intensity level in decibels:
Equation:

$$10\ \log_{10}\left(5.04 \times 10^{8}\right) = 10\ \left(8.70\right)\ \text{dB} = 87\ \text{dB}.$$

Discussion
This 87 dB sound has an intensity five times as great as an 80 dB sound. So a factor of five in intensity corresponds to a difference of 7 dB in sound intensity level. This value is true for any intensities differing by a factor of five.

Example:
Change Intensity Levels of a Sound: What Happens to the Decibel Level?
Show that if one sound is twice as intense as another, it has a sound level about 3 dB higher.
Strategy
You are given that the ratio of two intensities is 2 to 1, and are then asked to find the difference in their sound levels in decibels. You can solve this problem using of the properties of logarithms.
Solution
(1) Identify knowns:
The ratio of the two intensities is 2 to 1, or:
Equation:

$$\frac{I_2}{I_1} = 2.00.$$

We wish to show that the difference in sound levels is about 3 dB. That is, we want to show:
Equation:

$$\beta_2 - \beta_1 = 3\ \text{dB}.$$

Note that:
Equation:

$$\log_{10} b - \log_{10} a = \log_{10}\left(\frac{b}{a}\right).$$

(2) Use the definition of β to get:

Equation:

$$\beta_2 - \beta_1 = 10\ \log_{10}\left(\frac{I_2}{I_1}\right) = 10\ \log_{10}2.00 = 10\ (0.301)\ \text{dB}.$$

Thus,

Equation:

$$\beta_2 - \beta_1 = 3.01\ \text{dB}.$$

Discussion

This means that the two sound intensity levels differ by 3.01 dB, or about 3 dB, as advertised. Note that because only the ratio I_2/I_1 is given (and not the actual intensities), this result is true for any intensities that differ by a factor of two. For example, a 56.0 dB sound is twice as intense as a 53.0 dB sound, a 97.0 dB sound is half as intense as a 100 dB sound, and so on.

It should be noted at this point that there is another decibel scale in use, called the **sound pressure level**, based on the ratio of the pressure amplitude to a reference pressure. This scale is used particularly in applications where sound travels in water. It is beyond the scope of most introductory texts to treat this scale because it is not commonly used for sounds in air, but it is important to note that very different decibel levels may be encountered when sound pressure levels are quoted. For example, ocean noise pollution produced by ships may be as great as 200 dB expressed in the sound pressure level, where the more familiar sound intensity level we use here would be something under 140 dB for the same sound.

Note:

Take-Home Investigation: Feeling Sound

Find a CD player and a CD that has rock music. Place the player on a light table, insert the CD into the player, and start playing the CD. Place your hand gently on the table next to the speakers. Increase the volume and note the level when the table just begins to vibrate as the rock music plays. Increase the reading on the volume control until it doubles. What has happened to the vibrations?

Exercise:
Check Your Understanding

Problem:

Describe how amplitude is related to the loudness of a sound.

Solution:

Amplitude is directly proportional to the experience of loudness. As amplitude increases, loudness increases.

Exercise:
Check Your Understanding

Problem:

Identify common sounds at the levels of 10 dB, 50 dB, and 100 dB.

Solution:

10 dB: Running fingers through your hair.

50 dB: Inside a quiet home with no television or radio.

100 dB: Take-off of a jet plane.

Section Summary

- Intensity is the same for a sound wave as was defined for all waves; it is
  Equation:

  $$I = \frac{P}{A},$$

  where P is the power crossing area A. The SI unit for I is watts per meter squared. The intensity of a sound wave is also related to the pressure amplitude Δp
  Equation:

  $$I = \frac{(\Delta p)^2}{2\rho v_{\text{w}}},$$

  where ρ is the density of the medium in which the sound wave travels and v_{w} is the speed of sound in the medium.

- Sound intensity level in units of decibels (dB) is
  Equation:

  $$\beta\,(\text{dB}) = 10\,\log_{10}\left(\frac{I}{I_0}\right),$$

  where $I_0 = 10^{-12}\ \text{W/m}^2$ is the threshold intensity of hearing.

Conceptual Questions

Exercise:

Problem:

Six members of a synchronized swim team wear earplugs to protect themselves against water pressure at depths, but they can still hear the music and perform the combinations in the water perfectly. One day, they were asked to leave the pool so the dive team could practice a few dives, and they tried to practice on a mat, but seemed to have a lot more difficulty. Why might this be?

Exercise:

Problem:

A community is concerned about a plan to bring train service to their downtown from the town’s outskirts. The current sound intensity level, even though the rail yard is blocks away, is 70 dB downtown. The mayor assures the public that there will be a difference of only 30 dB in sound in the downtown area. Should the townspeople be concerned? Why?

Problems & Exercises

Exercise:

Problem:

What is the intensity in watts per meter squared of 85.0-dB sound?

Solution:
Equation:

$$3.16 \times 10^{-4}\ \mathrm{W/m^2}$$

Exercise:

Problem:

The warning tag on a lawn mower states that it produces noise at a level of 91.0 dB. What is this in watts per meter squared?

Exercise:

Problem:

A sound wave traveling in $20^{\circ}\mathrm{C}$ air has a pressure amplitude of 0.5 Pa. What is the intensity of the wave?

Solution:
Equation:

$$3.04 \times 10^{-4}\ \mathrm{W/m^2}$$

Exercise:

Problem:

What intensity level does the sound in the preceding problem correspond to?

Exercise:

Problem:

What sound intensity level in dB is produced by earphones that create an intensity of $4.00 \times 10^{-2}\ \mathrm{W/m^2}$?

Solution:

106 dB

Exercise:

Problem:

Show that an intensity of $10^{-12}\ \mathrm{W/m^2}$ is the same as $10^{-16}\ \mathrm{W/cm^2}$.

Exercise:

Problem:

(a) What is the decibel level of a sound that is twice as intense as a 90.0-dB sound? (b) What is the decibel level of a sound that is one-fifth as intense as a 90.0-dB sound?

Solution:

(a) 93 dB

(b) 83 dB

Exercise:

Problem:

(a) What is the intensity of a sound that has a level 7.00 dB lower than a $4.00 \times 10^{-9}\ \mathrm{W/m^2}$ sound? (b) What is the intensity of a sound that is 3.00 dB higher than a $4.00 \times 10^{-9}\ \mathrm{W/m^2}$ sound?

Exercise:

Problem:

(a) How much more intense is a sound that has a level 17.0 dB higher than another? (b) If one sound has a level 23.0 dB less than another, what is the ratio of their intensities?

Solution:

(a) 50.1

(b) 5.01×10^{-3} or $\frac{1}{200}$

Exercise:

Problem:

People with good hearing can perceive sounds as low in level as –8.00 dB at a frequency of 3000 Hz. What is the intensity of this sound in watts per meter squared?

Exercise:

Problem:

If a large housefly 3.0 m away from you makes a noise of 40.0 dB, what is the noise level of 1000 flies at that distance, assuming interference has a negligible effect?

Solution:

70.0 dB

Exercise:

Problem:

Ten cars in a circle at a boom box competition produce a 120-dB sound intensity level at the center of the circle. What is the average sound intensity level produced there by each stereo, assuming interference effects can be neglected?

Exercise:

Problem:

The amplitude of a sound wave is measured in terms of its maximum gauge pressure. By what factor does the amplitude of a sound wave increase if the sound intensity level goes up by 40.0 dB?

Solution:

100

Exercise:

Problem:

If a sound intensity level of 0 dB at 1000 Hz corresponds to a maximum gauge pressure (sound amplitude) of $10^{-9}\ \text{atm}$, what is the maximum gauge pressure in a 60-dB sound? What is the maximum gauge pressure in a 120-dB sound?

Exercise:

Problem:

An 8-hour exposure to a sound intensity level of 90.0 dB may cause hearing damage. What energy in joules falls on a 0.800-cm-diameter eardrum so exposed?

Solution:
Equation:

$$1.45 \times 10^{-3}\ \text{J}$$

Exercise:

Problem:

(a) Ear trumpets were never very common, but they did aid people with hearing losses by gathering sound over a large area and concentrating it on the smaller area of the eardrum. What decibel increase does an ear trumpet produce if its sound gathering area is $900\ \text{cm}^2$ and the area of the eardrum is $0.500\ \text{cm}^2$, but the trumpet only has an efficiency of 5.00% in transmitting the sound to the eardrum? (b) Comment on the usefulness of the decibel increase found in part (a).

Exercise:

Problem:

Sound is more effectively transmitted into a stethoscope by direct contact than through the air, and it is further intensified by being concentrated on the smaller area of the eardrum. It is reasonable to assume that sound is transmitted into a stethoscope 100 times as effectively compared with transmission though the air. What, then, is the gain in decibels produced by a stethoscope that has a sound gathering area of $15.0\ \text{cm}^2$, and concentrates the sound onto two eardrums with a total area of $0.900\ \text{cm}^2$ with an efficiency of 40.0%?

Solution:

28.2 dB

Exercise:

Problem:

Loudspeakers can produce intense sounds with surprisingly small energy input in spite of their low efficiencies. Calculate the power input needed to produce a 90.0-dB sound intensity level for a 12.0-cm-diameter speaker that has an efficiency of 1.00%. (This value is the sound intensity level right at the speaker.)

Glossary

intensity
: the power per unit area carried by a wave

sound intensity level
: a unitless quantity telling you the level of the sound relative to a fixed standard

sound pressure level
: the ratio of the pressure amplitude to a reference pressure

Doppler Effect and Sonic Booms

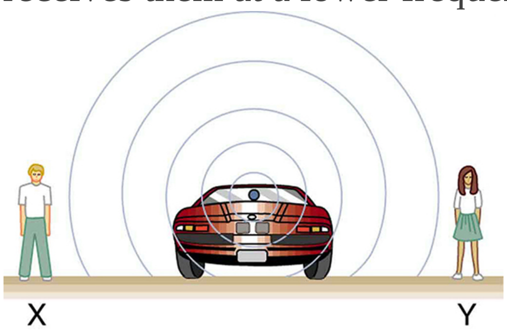
- Define Doppler effect, Doppler shift, and sonic boom.
- Calculate the frequency of a sound heard by someone observing Doppler shift.
- Describe the sounds produced by objects moving faster than the speed of sound.

The characteristic sound of a motorcycle buzzing by is an example of the **Doppler effect**. The high-pitch scream shifts dramatically to a lower-pitch roar as the motorcycle passes by a stationary observer. The closer the motorcycle brushes by, the more abrupt the shift. The faster the motorcycle moves, the greater the shift. We also hear this characteristic shift in frequency for passing race cars, airplanes, and trains. It is so familiar that it is used to imply motion and children often mimic it in play.

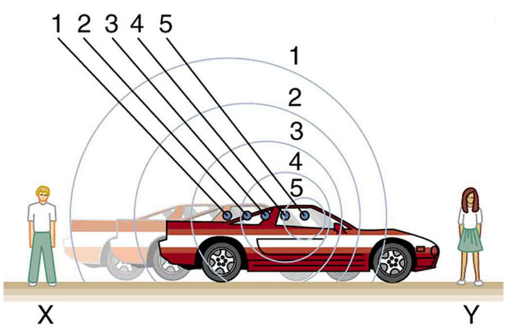
The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer. Although less familiar, this effect is easily noticed for a stationary source and moving observer. For example, if you ride a train past a stationary warning bell, you will hear the bell's frequency shift from high to low as you pass by. The actual change in frequency due to relative motion of source and observer is called a **Doppler shift**. The Doppler effect and Doppler shift are named for the Austrian physicist and mathematician Christian Johann Doppler (1803–1853), who did experiments with both moving sources and moving observers. Doppler, for example, had musicians play on a moving open train car and also play standing next to the train tracks as a train passed by. Their music was observed both on and off the train, and changes in frequency were measured.

What causes the Doppler shift? [link], [link], and [link] compare sound waves emitted by stationary and moving sources in a stationary air mass. Each disturbance spreads out spherically from the point where the sound was emitted. If the source is stationary, then all of the spheres representing the air compressions in the sound wave centered on the same point, and the stationary observers on either side see the same wavelength and frequency as emitted by the source, as in [link]. If the source is moving, as in [link], then the situation is different. Each compression of the air moves out in a

sphere from the point where it was emitted, but the point of emission moves. This moving emission point causes the air compressions to be closer together on one side and farther apart on the other. Thus, the wavelength is shorter in the direction the source is moving (on the right in [link]), and longer in the opposite direction (on the left in [link]). Finally, if the observers move, as in [link], the frequency at which they receive the compressions changes. The observer moving toward the source receives them at a higher frequency, and the person moving away from the source receives them at a lower frequency.

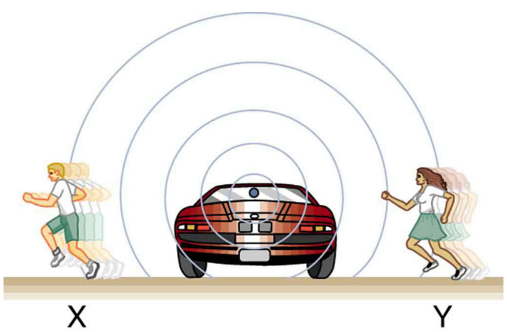


Sounds emitted by a source spread out in spherical waves. Because the source, observers, and air are stationary, the wavelength and frequency are the same in all directions and to all observers.



Sounds emitted by a

source moving to the right spread out from the points at which they were emitted. The wavelength is reduced and, consequently, the frequency is increased in the direction of motion, so that the observer on the right hears a higher-pitch sound. The opposite is true for the observer on the left, where the wavelength is increased and the frequency is reduced.



The same effect is produced when the observers move relative to the source. Motion toward the source increases frequency as the observer on the right passes through more wave crests than she would if stationary. Motion away from the

source decreases frequency as the observer on the left passes through fewer wave crests than he would if stationary.

We know that wavelength and frequency are related by $v_{\text{w}} = f\lambda$, where v_{w} is the fixed speed of sound. The sound moves in a medium and has the same speed v_{w} in that medium whether the source is moving or not. Thus f multiplied by λ is a constant. Because the observer on the right in [link] receives a shorter wavelength, the frequency she receives must be higher. Similarly, the observer on the left receives a longer wavelength, and hence he hears a lower frequency. The same thing happens in [link]. A higher frequency is received by the observer moving toward the source, and a lower frequency is received by an observer moving away from the source. In general, then, relative motion of source and observer toward one another increases the received frequency. Relative motion apart decreases frequency. The greater the relative speed is, the greater the effect.

Note:
The Doppler Effect
The Doppler effect occurs not only for sound but for any wave when there is relative motion between the observer and the source. There are Doppler shifts in the frequency of sound, light, and water waves, for example. Doppler shifts can be used to determine velocity, such as when ultrasound is reflected from blood in a medical diagnostic. The recession of galaxies is determined by the shift in the frequencies of light received from them and has implied much about the origins of the universe. Modern physics has been profoundly affected by observations of Doppler shifts.

For a stationary observer and a moving source, the frequency f_{obs} received by the observer can be shown to be

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right),$$

where f_{s} is the frequency of the source, v_{s} is the speed of the source along a line joining the source and observer, and v_{w} is the speed of sound. The minus sign is used for motion toward the observer and the plus sign for motion away from the observer, producing the appropriate shifts up and down in frequency. Note that the greater the speed of the source, the greater the effect. Similarly, for a stationary source and moving observer, the frequency received by the observer f_{obs} is given by

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right),$$

where v_{obs} is the speed of the observer along a line joining the source and observer. Here the plus sign is for motion toward the source, and the minus is for motion away from the source.

Example:

Calculate Doppler Shift: A Train Horn

Suppose a train that has a 150-Hz horn is moving at 35.0 m/s in still air on a day when the speed of sound is 340 m/s.

(a) What frequencies are observed by a stationary person at the side of the tracks as the train approaches and after it passes?

(b) What frequency is observed by the train’s engineer traveling on the train?

Strategy

To find the observed frequency in (a), $f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right)$, must be used because the source is moving. The minus sign is used for the approaching

train, and the plus sign for the receding train. In (b), there are two Doppler shifts—one for a moving source and the other for a moving observer.

Solution for (a)

(1) Enter known values into $f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} - v_{\text{s}}}\right)$.

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} - v_{\text{s}}}\right) = (150\text{ Hz})\left(\frac{340\text{ m/s}}{340\text{ m/s} - 35.0\text{ m/s}}\right)$$

(2) Calculate the frequency observed by a stationary person as the train approaches.

Equation:

$$f_{\text{obs}} = (150\text{ Hz})(1.11) = 167\text{ Hz}$$

(3) Use the same equation with the plus sign to find the frequency heard by a stationary person as the train recedes.

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} + v_{\text{s}}}\right) = (150\text{ Hz})\left(\frac{340\text{ m/s}}{340\text{ m/s} + 35.0\text{ m/s}}\right)$$

(4) Calculate the second frequency.

Equation:

$$f_{\text{obs}} = (150\text{ Hz})(0.907) = 136\text{ Hz}$$

Discussion on (a)

The numbers calculated are valid when the train is far enough away that the motion is nearly along the line joining train and observer. In both cases, the shift is significant and easily noticed. Note that the shift is 17.0 Hz for motion toward and 14.0 Hz for motion away. The shifts are not symmetric.

Solution for (b)

(1) Identify knowns:

- It seems reasonable that the engineer would receive the same frequency as emitted by the horn, because the relative velocity

between them is zero.
- Relative to the medium (air), the speeds are $v_s = v_{obs} = 35.0 \text{ m/s}$.
- The first Doppler shift is for the moving observer; the second is for the moving source.

(2) Use the following equation:
Equation:

$$f_{obs} = \left[f_s \left(\frac{v_w \pm v_{obs}}{v_w} \right) \right] \left(\frac{v_w}{v_w \pm v_s} \right).$$

The quantity in the square brackets is the Doppler-shifted frequency due to a moving observer. The factor on the right is the effect of the moving source.
(3) Because the train engineer is moving in the direction toward the horn, we must use the plus sign for v_{obs}; however, because the horn is also moving in the direction away from the engineer, we also use the plus sign for v_s. But the train is carrying both the engineer and the horn at the same velocity, so $v_s = v_{obs}$. As a result, everything but f_s cancels, yielding
Equation:

$$f_{obs} = f_s.$$

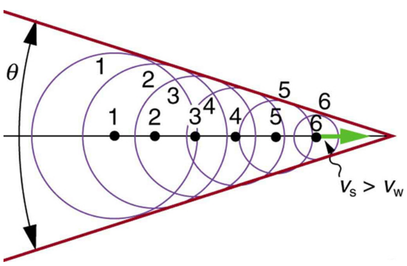
Discussion for (b)
We may expect that there is no change in frequency when source and observer move together because it fits your experience. For example, there is no Doppler shift in the frequency of conversations between driver and passenger on a motorcycle. People talking when a wind moves the air between them also observe no Doppler shift in their conversation. The crucial point is that source and observer are not moving relative to each other.

Sonic Booms to Bow Wakes

What happens to the sound produced by a moving source, such as a jet airplane, that approaches or even exceeds the speed of sound? The answer

to this question applies not only to sound but to all other waves as well.

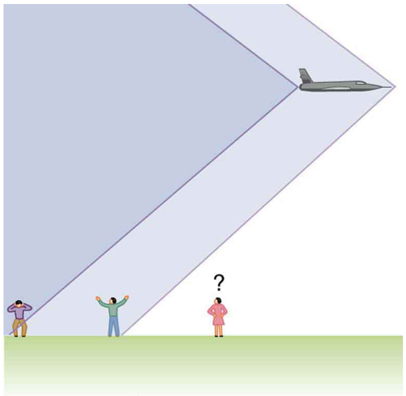
Suppose a jet airplane is coming nearly straight at you, emitting a sound of frequency f_s. The greater the plane's speed v_s, the greater the Doppler shift and the greater the value observed for f_{obs}. Now, as v_s approaches the speed of sound, f_{obs} approaches infinity, because the denominator in $f_{obs} = f_s\left(\frac{v_w}{v_w \pm v_s}\right)$ approaches zero. At the speed of sound, this result means that in front of the source, each successive wave is superimposed on the previous one because the source moves forward at the speed of sound. The observer gets them all at the same instant, and so the frequency is infinite. (Before airplanes exceeded the speed of sound, some people argued it would be impossible because such constructive superposition would produce pressures great enough to destroy the airplane.) If the source exceeds the speed of sound, no sound is received by the observer until the source has passed, so that the sounds from the approaching source are mixed with those from it when receding. This mixing appears messy, but something interesting happens—a sonic boom is created. (See [link].)



Sound waves from a source that moves faster than the speed of sound spread spherically from the point where they are emitted, but the source moves ahead of each.

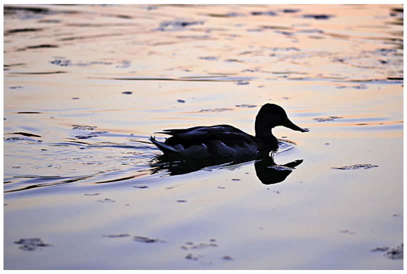
Constructive interference along the lines shown (actually a cone in three dimensions) creates a shock wave called a sonic boom. The faster the speed of the source, the smaller the angle θ.

There is constructive interference along the lines shown (a cone in three dimensions) from similar sound waves arriving there simultaneously. This superposition forms a disturbance called a **sonic boom**, a constructive interference of sound created by an object moving faster than sound. Inside the cone, the interference is mostly destructive, and so the sound intensity there is much less than on the shock wave. An aircraft creates two sonic booms, one from its nose and one from its tail. (See [link].) During television coverage of space shuttle landings, two distinct booms could often be heard. These were separated by exactly the time it would take the shuttle to pass by a point. Observers on the ground often do not see the aircraft creating the sonic boom, because it has passed by before the shock wave reaches them, as seen in [link]. If the aircraft flies close by at low altitude, pressures in the sonic boom can be destructive and break windows as well as rattle nerves. Because of how destructive sonic booms can be, supersonic flights are banned over populated areas of the United States.

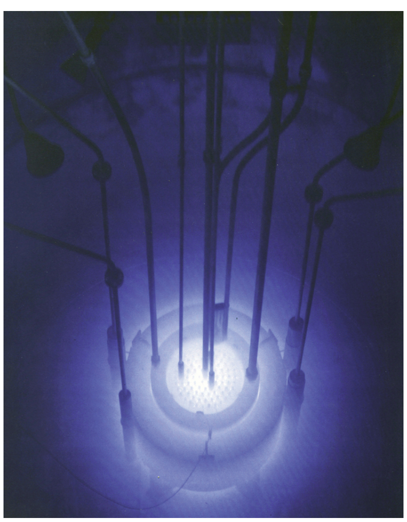


Two sonic booms, created by the nose and tail of an aircraft, are observed on the ground after the plane has passed by.

Sonic booms are one example of a broader phenomenon called bow wakes. A **bow wake**, such as the one in [link], is created when the wave source moves faster than the wave propagation speed. Water waves spread out in circles from the point where created, and the bow wake is the familiar V-shaped wake trailing the source. A more exotic bow wake is created when a subatomic particle travels through a medium faster than the speed of light travels in that medium. (In a vacuum, the maximum speed of light will be $c = 3.00 \times 10^8 \text{ m/s}$; in the medium of water, the speed of light is closer to $0.75c$. If the particle creates light in its passage, that light spreads on a cone with an angle indicative of the speed of the particle, as illustrated in [link]. Such a bow wake is called Cerenkov radiation and is commonly observed in particle physics.



Bow wake created by a duck. Constructive interference produces the rather structured wake, while there is relatively little wave action inside the wake, where interference is mostly destructive. (credit: Horia Varlan, Flickr)



The blue glow in this research reactor pool is Cerenkov radiation caused by subatomic particles traveling faster than the speed of light in water. (credit: U.S. Nuclear Regulatory Commission)

Doppler shifts and sonic booms are interesting sound phenomena that occur in all types of waves. They can be of considerable use. For example, the Doppler shift in ultrasound can be used to measure blood velocity, while police use the Doppler shift in radar (a microwave) to measure car velocities. In meteorology, the Doppler shift is used to track the motion of storm clouds; such “Doppler Radar” can give velocity and direction and rain or snow potential of imposing weather fronts. In astronomy, we can examine the light emitted from distant galaxies and determine their speed relative to ours. As galaxies move away from us, their light is shifted to a lower frequency, and so to a longer wavelength—the so-called red shift. Such information from galaxies far, far away has allowed us to estimate the age of the universe (from the Big Bang) as about 14 billion years.

Exercise:

Check Your Understanding

Problem:

Why did scientist Christian Doppler observe musicians both on a moving train and also from a stationary point not on the train?

Solution:

Doppler needed to compare the perception of sound when the observer is stationary and the sound source moves, as well as when the sound

source and the observer are both in motion.

Exercise:
Check Your Understanding

Problem:

Describe a situation in your life when you might rely on the Doppler shift to help you either while driving a car or walking near traffic.

Solution:

If I am driving and I hear Doppler shift in an ambulance siren, I would be able to tell when it was getting closer and also if it has passed by. This would help me to know whether I needed to pull over and let the ambulance through.

Section Summary

- The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer.
- The actual change in frequency is called the Doppler shift.
- A sonic boom is constructive interference of sound created by an object moving faster than sound.
- A sonic boom is a type of bow wake created when any wave source moves faster than the wave propagation speed.
- For a stationary observer and a moving source, the observed frequency f_{obs} is:
  Equation:
  $$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right),$$
  where f_{s} is the frequency of the source, v_{s} is the speed of the source, and v_{w} is the speed of sound. The minus sign is used for motion toward the observer and the plus sign for motion away.
- For a stationary source and moving observer, the observed frequency is:

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right),$$

where v_{obs} is the speed of the observer.

Conceptual Questions

Exercise:

Problem: Is the Doppler shift real or just a sensory illusion?

Exercise:

Problem:

Due to efficiency considerations related to its bow wake, the supersonic transport aircraft must maintain a cruising speed that is a constant ratio to the speed of sound (a constant Mach number). If the aircraft flies from warm air into colder air, should it increase or decrease its speed? Explain your answer.

Exercise:

Problem:

When you hear a sonic boom, you often cannot see the plane that made it. Why is that?

Problems & Exercises

Exercise:

Problem:

(a) What frequency is received by a person watching an oncoming ambulance moving at 110 km/h and emitting a steady 800-Hz sound from its siren? The speed of sound on this day is 345 m/s. (b) What frequency does she receive after the ambulance has passed?

Solution:

(a) 878 Hz

(b) 735 Hz

Exercise:

Problem:

(a) At an air show a jet flies directly toward the stands at a speed of 1200 km/h, emitting a frequency of 3500 Hz, on a day when the speed of sound is 342 m/s. What frequency is received by the observers? (b) What frequency do they receive as the plane flies directly away from them?

Exercise:

Problem:

What frequency is received by a mouse just before being dispatched by a hawk flying at it at 25.0 m/s and emitting a screech of frequency 3500 Hz? Take the speed of sound to be 331 m/s.

Solution:
Equation:

$$3.79 \times 10^{3} \text{ Hz}$$

Exercise:

Problem:

A spectator at a parade receives an 888-Hz tone from an oncoming trumpeter who is playing an 880-Hz note. At what speed is the musician approaching if the speed of sound is 338 m/s?

Exercise:

Problem:

A commuter train blows its 200-Hz horn as it approaches a crossing. The speed of sound is 335 m/s. (a) An observer waiting at the crossing receives a frequency of 208 Hz. What is the speed of the train? (b) What frequency does the observer receive as the train moves away?

Solution:

(a) 12.9 m/s

(b) 193 Hz

Exercise:

Problem:

Can you perceive the shift in frequency produced when you pull a tuning fork toward you at 10.0 m/s on a day when the speed of sound is 344 m/s? To answer this question, calculate the factor by which the frequency shifts and see if it is greater than 0.300%.

Exercise:

Problem:

Two eagles fly directly toward one another, the first at 15.0 m/s and the second at 20.0 m/s. Both screech, the first one emitting a frequency of 3200 Hz and the second one emitting a frequency of 3800 Hz. What frequencies do they receive if the speed of sound is 330 m/s?

Solution:

First eagle hears 4.23×10^3 Hz

Second eagle hears 3.56×10^3 Hz

Exercise:

Problem:

What is the minimum speed at which a source must travel toward you for you to be able to hear that its frequency is Doppler shifted? That is, what speed produces a shift of 0.300% on a day when the speed of sound is 331 m/s?

Glossary

Doppler effect
: an alteration in the observed frequency of a sound due to motion of either the source or the observer

Doppler shift
: the actual change in frequency due to relative motion of source and observer

sonic boom
: a constructive interference of sound created by an object moving faster than sound

bow wake
: V-shaped disturbance created when the wave source moves faster than the wave propagation speed

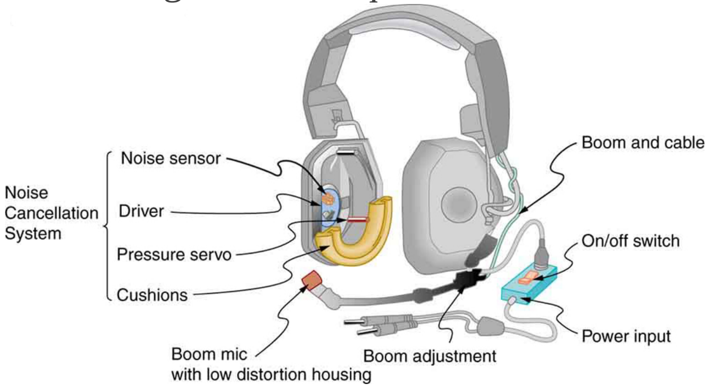
Sound Interference and Resonance: Standing Waves in Air Columns

- Define antinode, node, fundamental, overtones, and harmonics.
- Identify instances of sound interference in everyday situations.
- Describe how sound interference occurring inside open and closed tubes changes the characteristics of the sound, and how this applies to sounds produced by musical instruments.
- Calculate the length of a tube using sound wave measurements.

Some types of headphones use the phenomena of constructive and destructive interference to cancel out outside noises. (credit: JVC America, Flickr)

Interference is the hallmark of waves, all of which exhibit constructive and destructive interference exactly analogous to that seen for water waves. In fact, one way to prove something “is a wave” is to observe interference effects. So, sound being a wave, we expect it to exhibit interference; we have already mentioned a few such effects, such as the beats from two similar notes played simultaneously.

[link] shows a clever use of sound interference to cancel noise. Larger-scale applications of active noise reduction by destructive interference are contemplated for entire passenger compartments in commercial aircraft. To obtain destructive interference, a fast electronic analysis is performed, and a second sound is introduced with its maxima and minima exactly reversed from the incoming noise. Sound waves in fluids are pressure waves and consistent with Pascal’s principle; pressures from two different sources add and subtract like simple numbers; that is, positive and negative gauge pressures add to a much smaller pressure, producing a lower-intensity sound. Although completely destructive interference is possible only under the simplest conditions, it is possible to reduce noise levels by 30 dB or more using this technique.



Headphones designed to cancel noise with destructive interference create a sound wave exactly opposite to the incoming sound. These headphones can be more effective than the simple passive attenuation used in most ear protection. Such headphones were

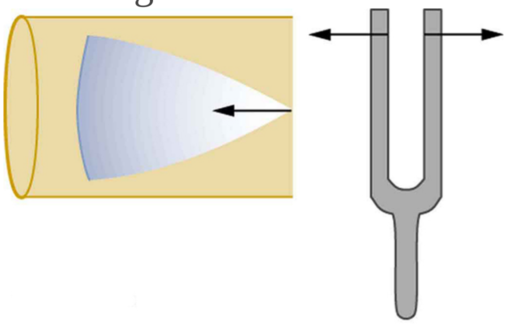
used on the record-setting, around the world nonstop flight of the Voyager aircraft to protect the pilots' hearing from engine noise.

Where else can we observe sound interference? All sound resonances, such as in musical instruments, are due to constructive and destructive interference. Only the resonant frequencies interfere constructively to form standing waves, while others interfere destructively and are absent. From the toot made by blowing over a bottle, to the characteristic flavor of a violin's sounding box, to the recognizability of a great singer's voice, resonance and standing waves play a vital role.

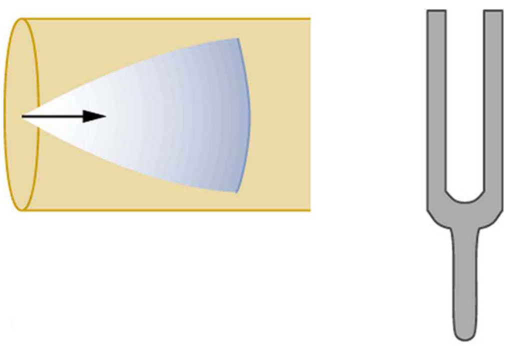
Note:
Interference
Interference is such a fundamental aspect of waves that observing interference is proof that something is a wave. The wave nature of light was established by experiments showing interference. Similarly, when electrons scattered from crystals exhibited interference, their wave nature was confirmed to be exactly as predicted by symmetry with certain wave characteristics of light.

Suppose we hold a tuning fork near the end of a tube that is closed at the other end, as shown in [link], [link], [link], and [link]. If the tuning fork has just the right frequency, the air column in the tube resonates loudly, but at most frequencies it vibrates very little. This observation just means that the air column has only certain natural frequencies. The figures show how a resonance at the lowest of these natural frequencies is formed. A disturbance travels down the tube at the speed of sound and bounces off the closed end. If the tube is just the right length, the reflected sound arrives back at the tuning fork exactly half a cycle later, and it interferes

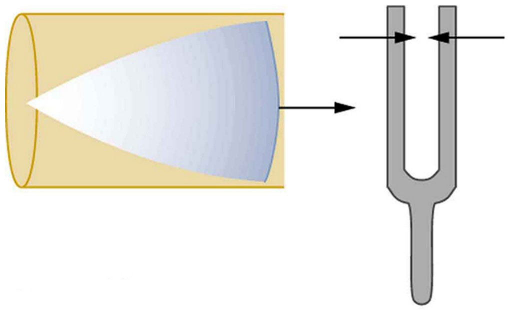
constructively with the continuing sound produced by the tuning fork. The incoming and reflected sounds form a standing wave in the tube as shown.



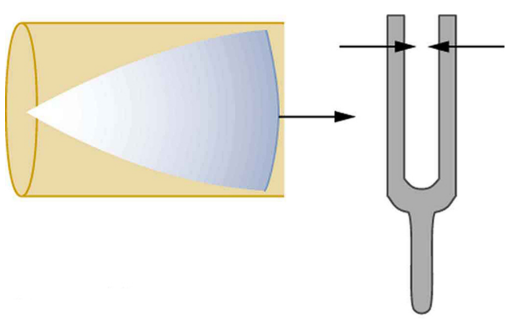
Resonance of air in a tube closed at one end, caused by a tuning fork. A disturbance moves down the tube.



Resonance of air in a tube closed at one end, caused by a tuning fork. The disturbance reflects from the closed end of the tube.



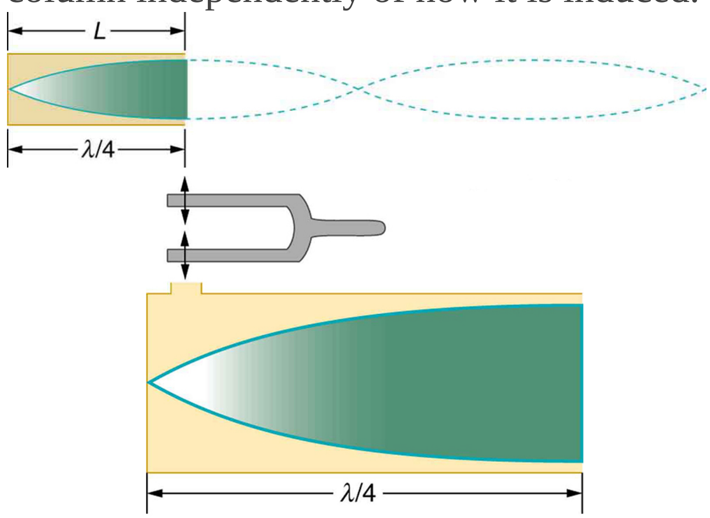
Resonance of air in a tube closed at one end, caused by a tuning fork. If the length of the tube L is just right, the disturbance gets back to the tuning fork half a cycle later and interferes constructively with the continuing sound from the tuning fork. This interference forms a standing wave, and the air column resonates.



Resonance of air in a tube closed at one end, caused by a tuning fork. A graph of air displacement along the length of the tube shows none at the closed

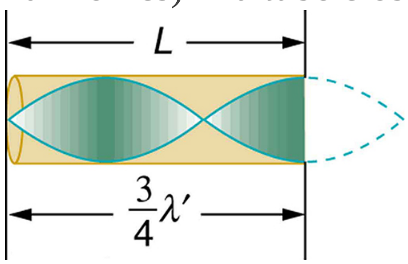
end, where the motion is constrained, and a maximum at the open end. This standing wave has one-fourth of its wavelength in the tube, so that $\lambda = 4L$.

The standing wave formed in the tube has its maximum air displacement (an **antinode**) at the open end, where motion is unconstrained, and no displacement (a **node**) at the closed end, where air movement is halted. The distance from a node to an antinode is one-fourth of a wavelength, and this equals the length of the tube; thus, $\lambda = 4L$. This same resonance can be produced by a vibration introduced at or near the closed end of the tube, as shown in [link]. It is best to consider this a natural vibration of the air column independently of how it is induced.

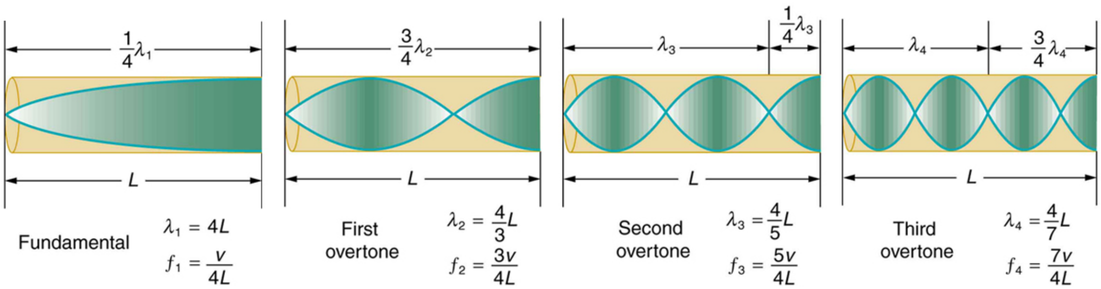


The same standing wave is created in the tube by a vibration introduced near its closed end.

Given that maximum air displacements are possible at the open end and none at the closed end, there are other, shorter wavelengths that can resonate in the tube, such as the one shown in [link]. Here the standing wave has three-fourths of its wavelength in the tube, or $L = (3/4)\lambda\prime$, so that $\lambda\prime = 4L/3$. Continuing this process reveals a whole series of shorter-wavelength and higher-frequency sounds that resonate in the tube. We use specific terms for the resonances in any system. The lowest resonant frequency is called the **fundamental**, while all higher resonant frequencies are called **overtones**. All resonant frequencies are integral multiples of the fundamental, and they are collectively called **harmonics**. The fundamental is the first harmonic, the first overtone is the second harmonic, and so on. [link] shows the fundamental and the first three overtones (the first four harmonics) in a tube closed at one end.

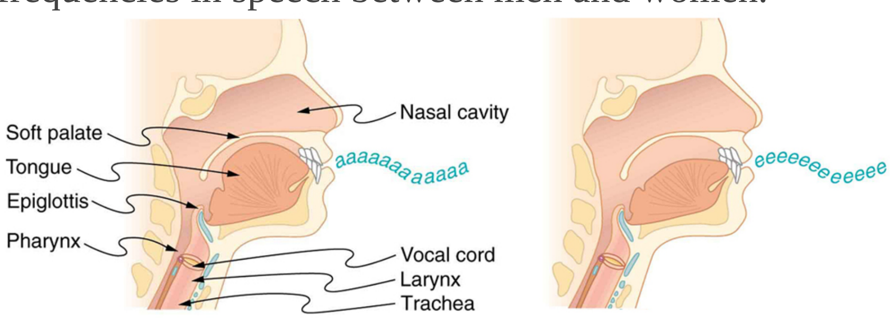


Another resonance for a tube closed at one end. This has maximum air displacements at the open end, and none at the closed end. The wavelength is shorter, with three-fourths $\lambda\prime$ equaling the length of the tube, so that $\lambda\prime = 4L/3$. This higher-frequency vibration is the first overtone.



The fundamental and three lowest overtones for a tube closed at one end. All have maximum air displacements at the open end and none at the closed end.

The fundamental and overtones can be present simultaneously in a variety of combinations. For example, middle C on a trumpet has a sound distinctively different from middle C on a clarinet, both instruments being modified versions of a tube closed at one end. The fundamental frequency is the same (and usually the most intense), but the overtones and their mix of intensities are different and subject to shading by the musician. This mix is what gives various musical instruments (and human voices) their distinctive characteristics, whether they have air columns, strings, sounding boxes, or drumheads. In fact, much of our speech is determined by shaping the cavity formed by the throat and mouth and positioning the tongue to adjust the fundamental and combination of overtones. Simple resonant cavities can be made to resonate with the sound of the vowels, for example. (See [link].) In boys, at puberty, the larynx grows and the shape of the resonant cavity changes giving rise to the difference in predominant frequencies in speech between men and women.



The throat and mouth form an air column closed at one end that resonates in response to vibrations in the voice box. The spectrum of overtones and their intensities vary with mouth shaping and tongue position to form different sounds. The voice box can be replaced with a mechanical vibrator, and understandable speech is still possible. Variations in basic shapes make different voices recognizable.

Now let us look for a pattern in the resonant frequencies for a simple tube that is closed at one end. The fundamental has $\lambda = 4L$, and frequency is related to wavelength and the speed of sound as given by:

Equation:

$$v_{\mathrm{w}} = f\lambda.$$

Solving for f in this equation gives

Equation:

$$f = \frac{v_{\mathrm{w}}}{\lambda} = \frac{v_{\mathrm{w}}}{4L},$$

where v_{w} is the speed of sound in air. Similarly, the first overtone has $\lambda\prime = 4L/3$ (see [link]), so that

Equation:

$$f\prime = 3\frac{v_{\mathrm{w}}}{4L} = 3f.$$

Because $f\prime = 3f$, we call the first overtone the third harmonic. Continuing this process, we see a pattern that can be generalized in a single expression. The resonant frequencies of a tube closed at one end are

Equation:

$$f_n = n\frac{v_\mathrm{w}}{4L}, n = 1,3,5,$$

where f_1 is the fundamental, f_3 is the first overtone, and so on. It is interesting that the resonant frequencies depend on the speed of sound and, hence, on temperature. This dependence poses a noticeable problem for organs in old unheated cathedrals, and it is also the reason why musicians commonly bring their wind instruments to room temperature before playing them.

Example:
Find the Length of a Tube with a 128 Hz Fundamental
(a) What length should a tube closed at one end have on a day when the air temperature, is $22.0^\mathrm{o}\mathrm{C}$, if its fundamental frequency is to be 128 Hz (C below middle C)?
(b) What is the frequency of its fourth overtone?
Strategy
The length L can be found from the relationship in $f_n = n\frac{v_\mathrm{w}}{4L}$, but we will first need to find the speed of sound v_w.
Solution for (a)
(1) Identify knowns:

- the fundamental frequency is 128 Hz
- the air temperature is $22.0^\mathrm{o}\mathrm{C}$

(2) Use $f_n = n\frac{v_\mathrm{w}}{4L}$ to find the fundamental frequency ($n = 1$).
Equation:

$$f_1 = \frac{v_\mathrm{w}}{4L}$$

(3) Solve this equation for length.
Equation:

$$L = \frac{v_\mathrm{w}}{4f_1}$$

(4) Find the speed of sound using $v_{\mathrm{w}} = (331\ \mathrm{m/s})\sqrt{\frac{T}{273\ \mathrm{K}}}$.

Equation:

$$v_{\mathrm{w}} = (331\ \mathrm{m/s})\sqrt{\frac{295\ \mathrm{K}}{273\ \mathrm{K}}} = 344\ \mathrm{m/s}$$

(5) Enter the values of the speed of sound and frequency into the expression for L.

Equation:

$$L = \frac{v_{\mathrm{w}}}{4f_1} = \frac{344\ \mathrm{m/s}}{4(128\ \mathrm{Hz})} = 0.672\ \mathrm{m}$$

Discussion on (a)

Many wind instruments are modified tubes that have finger holes, valves, and other devices for changing the length of the resonating air column and hence, the frequency of the note played. Horns producing very low frequencies, such as tubas, require tubes so long that they are coiled into loops.

Solution for (b)

(1) Identify knowns:

- the first overtone has $n = 3$
- the second overtone has $n = 5$
- the third overtone has $n = 7$
- the fourth overtone has $n = 9$

(2) Enter the value for the fourth overtone into $f_n = n\frac{v_{\mathrm{w}}}{4L}$.

Equation:

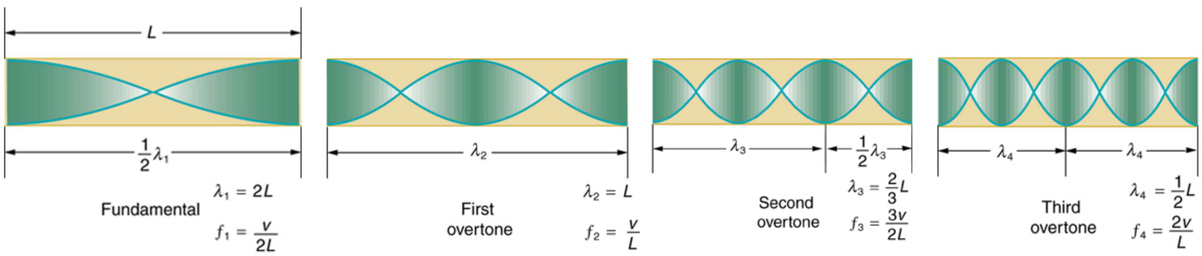
$$f_9 = 9\frac{v_{\mathrm{w}}}{4L} = 9f_1 = 1.15\ \mathrm{kHz}$$

Discussion on (b)

Whether this overtone occurs in a simple tube or a musical instrument depends on how it is stimulated to vibrate and the details of its shape. The

trombone, for example, does not produce its fundamental frequency and only makes overtones.

Another type of tube is one that is *open* at both ends. Examples are some organ pipes, flutes, and oboes. The resonances of tubes open at both ends can be analyzed in a very similar fashion to those for tubes closed at one end. The air columns in tubes open at both ends have maximum air displacements at both ends, as illustrated in [link]. Standing waves form as shown.



The resonant frequencies of a tube open at both ends are shown, including the fundamental and the first three overtones. In all cases the maximum air displacements occur at both ends of the tube, giving it different natural frequencies than a tube closed at one end.

Based on the fact that a tube open at both ends has maximum air displacements at both ends, and using [link] as a guide, we can see that the resonant frequencies of a tube open at both ends are:

Equation:

$$f_n = n\frac{v_{\text{w}}}{2L},\ n = 1, 2, 3...,$$

where f_1 is the fundamental, f_2 is the first overtone, f_3 is the second overtone, and so on. Note that a tube open at both ends has a fundamental frequency twice what it would have if closed at one end. It also has a different spectrum of overtones than a tube closed at one end. So if you had

two tubes with the same fundamental frequency but one was open at both ends and the other was closed at one end, they would sound different when played because they have different overtones. Middle C, for example, would sound richer played on an open tube, because it has even multiples of the fundamental as well as odd. A closed tube has only odd multiples.

Note:

Real-World Applications: Resonance in Everyday Systems

Resonance occurs in many different systems, including strings, air columns, and atoms. Resonance is the driven or forced oscillation of a system at its natural frequency. At resonance, energy is transferred rapidly to the oscillating system, and the amplitude of its oscillations grows until the system can no longer be described by Hooke's law. An example of this is the distorted sound intentionally produced in certain types of rock music.

Wind instruments use resonance in air columns to amplify tones made by lips or vibrating reeds. Other instruments also use air resonance in clever ways to amplify sound. [link] shows a violin and a guitar, both of which have sounding boxes but with different shapes, resulting in different overtone structures. The vibrating string creates a sound that resonates in the sounding box, greatly amplifying the sound and creating overtones that give the instrument its characteristic flavor. The more complex the shape of the sounding box, the greater its ability to resonate over a wide range of frequencies. The marimba, like the one shown in [link] uses pots or gourds below the wooden slats to amplify their tones. The resonance of the pot can be adjusted by adding water.

String instruments such as violins and guitars use resonance in their sounding boxes to amplify and enrich the sound created by their vibrating strings. The bridge and supports couple the string vibrations to the sounding boxes and air within. (credits: guitar, Feliciano Guimares, Fotopedia; violin, Steve Snodgrass, Flickr)

Resonance has been used in musical instruments since prehistoric times. This marimba uses gourds as resonance chambers to amplify its sound. (credit: APC Events, Flickr)

We have emphasized sound applications in our discussions of resonance and standing waves, but these ideas apply to any system that has wave characteristics. Vibrating strings, for example, are actually resonating and have fundamentals and overtones similar to those for air columns. More subtle are the resonances in atoms due to the wave character of their electrons. Their orbitals can be viewed as standing waves, which have a fundamental (ground state) and overtones (excited states). It is fascinating that wave characteristics apply to such a wide range of physical systems.

Exercise:

Check Your Understanding

Problem:

Describe how noise-canceling headphones differ from standard headphones used to block outside sounds.

Solution:

Regular headphones only block sound waves with a physical barrier. Noise-canceling headphones use destructive interference to reduce the loudness of outside sounds.

Exercise:
Check Your Understanding

Problem:

How is it possible to use a standing wave's node and antinode to determine the length of a closed-end tube?

Solution:

When the tube resonates at its natural frequency, the wave's node is located at the closed end of the tube, and the antinode is located at the open end. The length of the tube is equal to one-fourth of the wavelength of this wave. Thus, if we know the wavelength of the wave, we can determine the length of the tube.

Note:
PhET Explorations: Sound
This simulation lets you see sound waves. Adjust the frequency or volume and you can see and hear how the wave changes. Move the listener around and hear what she hears.
https://archive.cnx.org/specials/c4d3b96e-41f3-11e5-ab7b-47e22dffc18e/sound/#sim-single-source

Section Summary

- Sound interference and resonance have the same properties as defined for all waves.
- In air columns, the lowest-frequency resonance is called the fundamental, whereas all higher resonant frequencies are called overtones. Collectively, they are called harmonics.

- The resonant frequencies of a tube closed at one end are:
  Equation:

$$f_n = n\frac{v_{\mathrm{w}}}{4L}, n = 1, 3, 5...,$$

  f_1 is the fundamental and L is the length of the tube.
- The resonant frequencies of a tube open at both ends are:
  Equation:

$$f_n = n\frac{v_{\mathrm{w}}}{2L}, n = 1, 2, 3...$$

Conceptual Questions

Exercise:

Problem:

How does an unamplified guitar produce sounds so much more intense than those of a plucked string held taut by a simple stick?

Exercise:

Problem:

You are given two wind instruments of identical length. One is open at both ends, whereas the other is closed at one end. Which is able to produce the lowest frequency?

Exercise:

Problem:

What is the difference between an overtone and a harmonic? Are all harmonics overtones? Are all overtones harmonics?

Problems & Exercises

Exercise:

Problem:

A “showy” custom-built car has two brass horns that are supposed to produce the same frequency but actually emit 263.8 and 264.5 Hz. What beat frequency is produced?

Solution:

0.7 Hz

Exercise:

Problem:

What beat frequencies will be present: (a) If the musical notes A and C are played together (frequencies of 220 and 264 Hz)? (b) If D and F are played together (frequencies of 297 and 352 Hz)? (c) If all four are played together?

Exercise:

Problem:

What beat frequencies result if a piano hammer hits three strings that emit frequencies of 127.8, 128.1, and 128.3 Hz?

Solution:

0.3 Hz, 0.2 Hz, 0.5 Hz

Exercise:

Problem:

A piano tuner hears a beat every 2.00 s when listening to a 264.0-Hz tuning fork and a single piano string. What are the two possible frequencies of the string?

Exercise:

Problem:

(a) What is the fundamental frequency of a 0.672-m-long tube, open at both ends, on a day when the speed of sound is 344 m/s? (b) What is the frequency of its second harmonic?

Solution:

(a) 256 Hz

(b) 512 Hz

Exercise:

Problem:

If a wind instrument, such as a tuba, has a fundamental frequency of 32.0 Hz, what are its first three overtones? It is closed at one end. (The overtones of a real tuba are more complex than this example, because it is a tapered tube.)

Exercise:

Problem:

What are the first three overtones of a bassoon that has a fundamental frequency of 90.0 Hz? It is open at both ends. (The overtones of a real bassoon are more complex than this example, because its double reed makes it act more like a tube closed at one end.)

Solution:

180 Hz, 270 Hz, 360 Hz

Exercise:

Problem:

How long must a flute be in order to have a fundamental frequency of 262 Hz (this frequency corresponds to middle C on the evenly tempered chromatic scale) on a day when air temperature is 20.0°C? It is open at both ends.

Exercise:

Problem:

What length should an oboe have to produce a fundamental frequency of 110 Hz on a day when the speed of sound is 343 m/s? It is open at both ends.

Solution:

1.56 m

Exercise:

Problem:

What is the length of a tube that has a fundamental frequency of 176 Hz and a first overtone of 352 Hz if the speed of sound is 343 m/s?

Exercise:

Problem:

(a) Find the length of an organ pipe closed at one end that produces a fundamental frequency of 256 Hz when air temperature is 18.0°C. (b) What is its fundamental frequency at 25.0°C?

Solution:

(a) 0.334 m

(b) 259 Hz

Exercise:

Problem:

By what fraction will the frequencies produced by a wind instrument change when air temperature goes from $10.0^{\circ}C$ to $30.0^{\circ}C$? That is, find the ratio of the frequencies at those temperatures.

Exercise:

Problem:

The ear canal resonates like a tube closed at one end. (See [link].) If ear canals range in length from 1.80 to 2.60 cm in an average population, what is the range of fundamental resonant frequencies? Take air temperature to be $37.0^{\circ}C$, which is the same as body temperature. How does this result correlate with the intensity versus frequency graph ([link] of the human ear?

Solution:

3.39 to 4.90 kHz

Exercise:

Problem:

Calculate the first overtone in an ear canal, which resonates like a 2.40-cm-long tube closed at one end, by taking air temperature to be $37.0^{\circ}C$. Is the ear particularly sensitive to such a frequency? (The resonances of the ear canal are complicated by its nonuniform shape, which we shall ignore.)

Exercise:

Problem:

A crude approximation of voice production is to consider the breathing passages and mouth to be a resonating tube closed at one end. (See [link].) (a) What is the fundamental frequency if the tube is 0.240-m long, by taking air temperature to be $37.0^{\circ}C$? (b) What would this frequency become if the person replaced the air with helium? Assume the same temperature dependence for helium as for air.

Solution:

(a) 367 Hz

(b) 1.07 kHz

Exercise:

Problem:

(a) Students in a physics lab are asked to find the length of an air column in a tube closed at one end that has a fundamental frequency of 256 Hz. They hold the tube vertically and fill it with water to the top, then lower the water while a 256-Hz tuning fork is rung and listen for the first resonance. What is the air temperature if the resonance occurs for a length of 0.336 m? (b) At what length will they observe the second resonance (first overtone)?

Exercise:

Problem:

What frequencies will a 1.80-m-long tube produce in the audible range at 20.0°C if: (a) The tube is closed at one end? (b) It is open at both ends?

Solution:

(a) $f_n = n(47.6\text{ Hz}),\ n = 1, 3, 5, ..., 419$

(b) $f_n = n(95.3\text{ Hz}),\ n = 1, 2, 3, ..., 210$

Glossary

antinode
: point of maximum displacement

node

point of zero displacement

fundamental
: the lowest-frequency resonance

overtones
: all resonant frequencies higher than the fundamental

harmonics
: the term used to refer collectively to the fundamental and its overtones

Hearing

- Define hearing, pitch, loudness, timbre, note, tone, phon, ultrasound, and infrasound.
- Compare loudness to frequency and intensity of a sound.
- Identify structures of the inner ear and explain how they relate to sound perception.

Hearing allows this vocalist, his band, and his fans to enjoy music. (credit: West Point Public Affairs, Flickr)

The human ear has a tremendous range and sensitivity. It can give us a wealth of simple information—such as pitch, loudness, and direction. And from its input we can detect musical quality and nuances of voiced emotion. How is our hearing related to the physical qualities of sound, and how does the hearing mechanism work?

Hearing is the perception of sound. (Perception is commonly defined to be awareness through the senses, a typically circular definition of higher-level processes in living organisms.) Normal human hearing encompasses frequencies from 20 to 20,000 Hz, an impressive range. Sounds below 20 Hz are called **infrasound**, whereas those above 20,000 Hz are **ultrasound**. Neither is perceived by the ear, although infrasound can sometimes be felt as vibrations. When we do hear low-frequency vibrations, such as the

sounds of a diving board, we hear the individual vibrations only because there are higher-frequency sounds in each. Other animals have hearing ranges different from that of humans. Dogs can hear sounds as high as 30,000 Hz, whereas bats and dolphins can hear up to 100,000-Hz sounds. You may have noticed that dogs respond to the sound of a dog whistle which produces sound out of the range of human hearing. Elephants are known to respond to frequencies below 20 Hz.

The perception of frequency is called **pitch**. Most of us have excellent relative pitch, which means that we can tell whether one sound has a different frequency from another. Typically, we can discriminate between two sounds if their frequencies differ by 0.3% or more. For example, 500.0 and 501.5 Hz are noticeably different. Pitch perception is directly related to frequency and is not greatly affected by other physical quantities such as intensity. Musical **notes** are particular sounds that can be produced by most instruments and in Western music have particular names. Combinations of notes constitute music. Some people can identify musical notes, such as A-sharp, C, or E-flat, just by listening to them. This uncommon ability is called perfect pitch.

The ear is remarkably sensitive to low-intensity sounds. The lowest audible intensity or threshold is about $10^{-12}\ \mathrm{W/m^2}$ or 0 dB. Sounds as much as 10^{12} more intense can be briefly tolerated. Very few measuring devices are capable of observations over a range of a trillion. The perception of intensity is called **loudness**. At a given frequency, it is possible to discern differences of about 1 dB, and a change of 3 dB is easily noticed. But loudness is not related to intensity alone. Frequency has a major effect on how loud a sound seems. The ear has its maximum sensitivity to frequencies in the range of 2000 to 5000 Hz, so that sounds in this range are perceived as being louder than, say, those at 500 or 10,000 Hz, even when they all have the same intensity. Sounds near the high- and low-frequency extremes of the hearing range seem even less loud, because the ear is even less sensitive at those frequencies. [link] gives the dependence of certain human hearing perceptions on physical quantities.

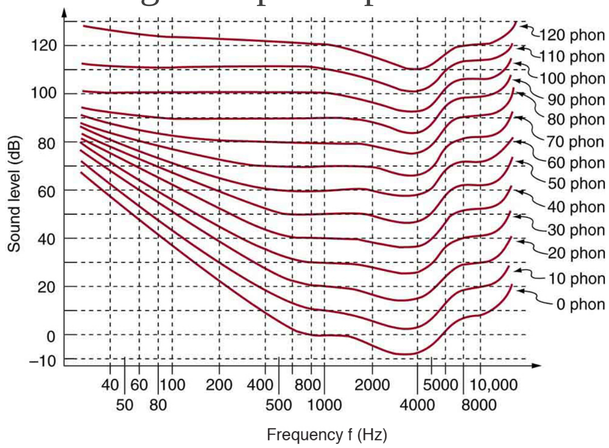
Perception	Physical quantity
Pitch	Frequency
Loudness	Intensity and Frequency
Timbre	Number and relative intensity of multiple frequencies. Subtle craftsmanship leads to non-linear effects and more detail.
Note	Basic unit of music with specific names, combined to generate tunes
Tone	Number and relative intensity of multiple frequencies.

Sound Perceptions

When a violin plays middle C, there is no mistaking it for a piano playing the same note. The reason is that each instrument produces a distinctive set of frequencies and intensities. We call our perception of these combinations of frequencies and intensities **tone** quality, or more commonly the **timbre** of the sound. It is more difficult to correlate timbre perception to physical quantities than it is for loudness or pitch perception. Timbre is more subjective. Terms such as dull, brilliant, warm, cold, pure, and rich are employed to describe the timbre of a sound. So the consideration of timbre takes us into the realm of perceptual psychology, where higher-level processes in the brain are dominant. This is true for other perceptions of sound, such as music and noise. We shall not delve further into them; rather, we will concentrate on the question of loudness perception.

A unit called a **phon** is used to express loudness numerically. Phons differ from decibels because the phon is a unit of loudness perception, whereas the decibel is a unit of physical intensity. [link] shows the relationship of loudness to intensity (or intensity level) and frequency for persons with normal hearing. The curved lines are equal-loudness curves. Each curve is

labeled with its loudness in phons. Any sound along a given curve will be perceived as equally loud by the average person. The curves were determined by having large numbers of people compare the loudness of sounds at different frequencies and sound intensity levels. At a frequency of 1000 Hz, phons are taken to be numerically equal to decibels. The following example helps illustrate how to use the graph:



The relationship of loudness in phons to intensity level (in decibels) and intensity (in watts per meter squared) for persons with normal hearing. The curved lines are equal-loudness curves—all sounds on a given curve are perceived as equally loud. Phons and decibels are defined to be the same at 1000 Hz.

Example:
Measuring Loudness: Loudness Versus Intensity Level and Frequency
(a) What is the loudness in phons of a 100-Hz sound that has an intensity level of 80 dB? (b) What is the intensity level in decibels of a 4000-Hz

sound having a loudness of 70 phons? (c) At what intensity level will an 8000-Hz sound have the same loudness as a 200-Hz sound at 60 dB?

Strategy for (a)

The graph in [link] should be referenced in order to solve this example. To find the loudness of a given sound, you must know its frequency and intensity level and locate that point on the square grid, then interpolate between loudness curves to get the loudness in phons.

Solution for (a)

(1) Identify knowns:

- The square grid of the graph relating phons and decibels is a plot of intensity level versus frequency—both physical quantities.
- 100 Hz at 80 dB lies halfway between the curves marked 70 and 80 phons.

(2) Find the loudness: 75 phons.

Strategy for (b)

The graph in [link] should be referenced in order to solve this example. To find the intensity level of a sound, you must have its frequency and loudness. Once that point is located, the intensity level can be determined from the vertical axis.

Solution for (b)

(1) Identify knowns:

- Values are given to be 4000 Hz at 70 phons.

(2) Follow the 70-phon curve until it reaches 4000 Hz. At that point, it is below the 70 dB line at about 67 dB.

(3) Find the intensity level:

67 dB

Strategy for (c)

The graph in [link] should be referenced in order to solve this example.

Solution for (c)

(1) Locate the point for a 200 Hz and 60 dB sound.

(2) Find the loudness: This point lies just slightly above the 50-phon curve, and so its loudness is 51 phons.

(3) Look for the 51-phon level is at 8000 Hz: 63 dB.

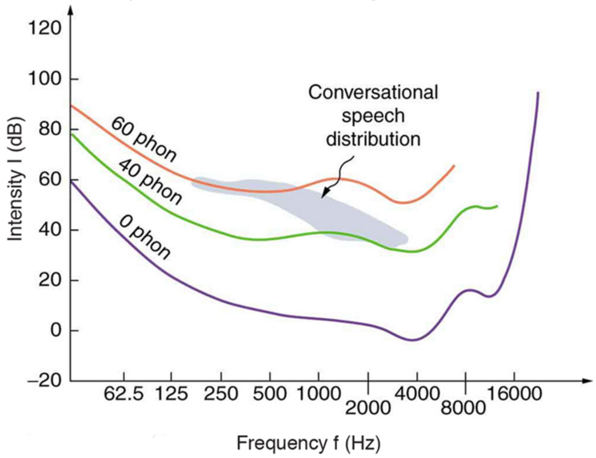
Discussion

These answers, like all information extracted from [link], have uncertainties of several phons or several decibels, partly due to difficulties in interpolation, but mostly related to uncertainties in the equal-loudness curves.

Further examination of the graph in [link] reveals some interesting facts about human hearing. First, sounds below the 0-phon curve are not perceived by most people. So, for example, a 60 Hz sound at 40 dB is inaudible. The 0-phon curve represents the threshold of normal hearing. We can hear some sounds at intensity levels below 0 dB. For example, a 3-dB, 5000-Hz sound is audible, because it lies above the 0-phon curve. The loudness curves all have dips in them between about 2000 and 5000 Hz. These dips mean the ear is most sensitive to frequencies in that range. For example, a 15-dB sound at 4000 Hz has a loudness of 20 phons, the same as a 20-dB sound at 1000 Hz. The curves rise at both extremes of the frequency range, indicating that a greater-intensity level sound is needed at those frequencies to be perceived to be as loud as at middle frequencies. For example, a sound at 10,000 Hz must have an intensity level of 30 dB to seem as loud as a 20 dB sound at 1000 Hz. Sounds above 120 phons are painful as well as damaging.

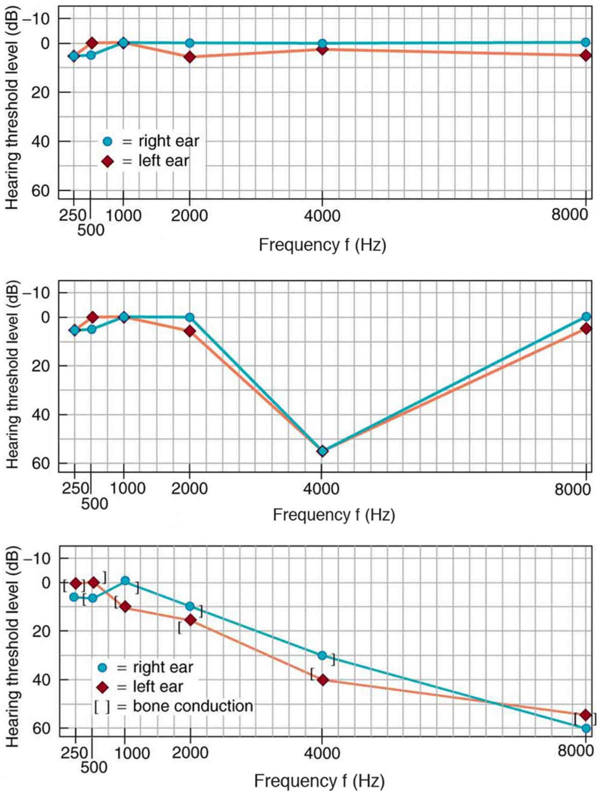
We do not often utilize our full range of hearing. This is particularly true for frequencies above 8000 Hz, which are rare in the environment and are unnecessary for understanding conversation or appreciating music. In fact, people who have lost the ability to hear such high frequencies are usually unaware of their loss until tested. The shaded region in [link] is the frequency and intensity region where most conversational sounds fall. The curved lines indicate what effect hearing losses of 40 and 60 phons will have. A 40-phon hearing loss at all frequencies still allows a person to understand conversation, although it will seem very quiet. A person with a 60-phon loss at all frequencies will hear only the lowest frequencies and will not be able to understand speech unless it is much louder than normal. Even so, speech may seem indistinct, because higher frequencies are not as well perceived. The conversational speech region also has a gender component, in that female voices are usually characterized by higher

frequencies. So the person with a 60-phon hearing impediment might have difficulty understanding the normal conversation of a woman.



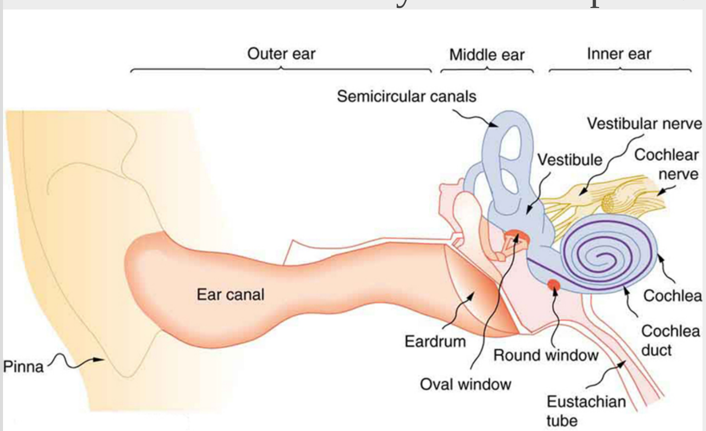
The shaded region represents frequencies and intensity levels found in normal conversational speech. The 0-phon line represents the normal hearing threshold, while those at 40 and 60 represent thresholds for people with 40- and 60-phon hearing losses, respectively.

Hearing tests are performed over a range of frequencies, usually from 250 to 8000 Hz, and can be displayed graphically in an audiogram like that in [link]. The hearing threshold is measured in dB *relative to the normal threshold*, so that normal hearing registers as 0 dB at all frequencies. Hearing loss caused by noise typically shows a dip near the 4000 Hz frequency, irrespective of the frequency that caused the loss and often affects both ears. The most common form of hearing loss comes with age and is called *presbycusis*—literally elder ear. Such loss is increasingly severe at higher frequencies, and interferes with music appreciation and speech recognition.



Audiograms showing the threshold in intensity level versus frequency for three different individuals. Intensity level is measured relative to the normal threshold. The top left graph is that of a person with normal hearing. The graph to its right has a dip at 4000 Hz and is that of a child who suffered hearing loss due to a cap gun. The third graph is typical of presbycusis, the progressive loss of higher frequency hearing with age. Tests performed by bone conduction (brackets) can distinguish nerve damage from middle ear damage.

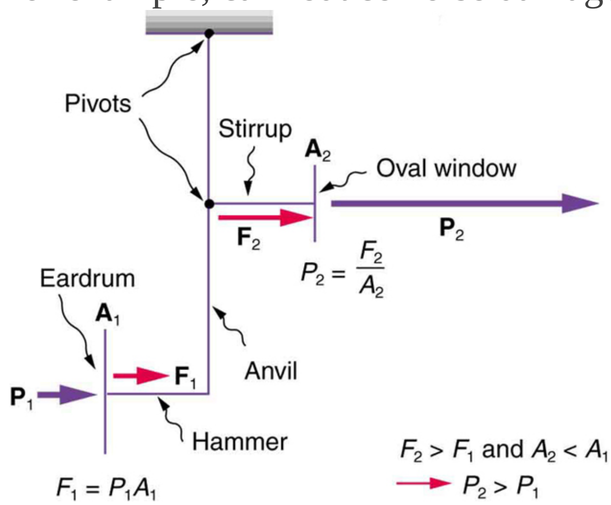
Note:
The Hearing Mechanism
The hearing mechanism involves some interesting physics. The sound wave that impinges upon our ear is a pressure wave. The ear is a transducer that converts sound waves into electrical nerve impulses in a manner much more sophisticated than, but analogous to, a microphone. [link] shows the gross anatomy of the ear with its division into three parts: the outer ear or ear canal; the middle ear, which runs from the eardrum to the cochlea; and the inner ear, which is the cochlea itself. The body part normally referred to as the ear is technically called the pinna.



The illustration shows the gross anatomy of the human ear.

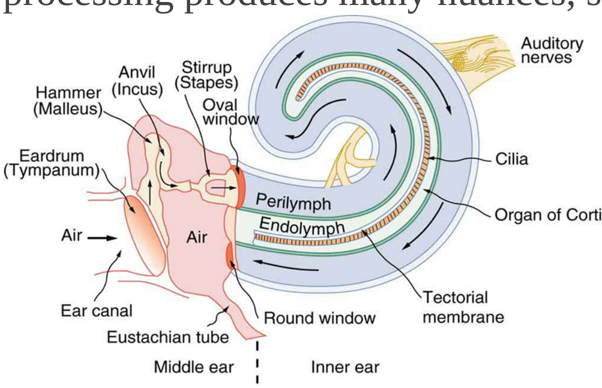
The outer ear, or ear canal, carries sound to the recessed protected eardrum. The air column in the ear canal resonates and is partially responsible for the sensitivity of the ear to sounds in the 2000 to 5000 Hz range. The middle ear converts sound into mechanical vibrations and applies these vibrations to the cochlea. The lever system of the middle ear takes the force exerted on the eardrum by sound pressure variations, amplifies it and transmits it to the

inner ear via the oval window, creating pressure waves in the cochlea approximately 40 times greater than those impinging on the eardrum. (See [link].) Two muscles in the middle ear (not shown) protect the inner ear from very intense sounds. They react to intense sound in a few milliseconds and reduce the force transmitted to the cochlea. This protective reaction can also be triggered by your own voice, so that humming while shooting a gun, for example, can reduce noise damage.



This schematic shows the middle ear's system for converting sound pressure into force, increasing that force through a lever system, and applying the increased force to a small area of the cochlea, thereby creating a pressure about 40 times that in the original sound wave. A protective muscle reaction to intense sounds greatly reduces the mechanical advantage of the lever system.

[link] shows the middle and inner ear in greater detail. Pressure waves moving through the cochlea cause the tectorial membrane to vibrate, rubbing cilia (called hair cells), which stimulate nerves that send electrical signals to the brain. The membrane resonates at different positions for different frequencies, with high frequencies stimulating nerves at the near end and low frequencies at the far end. The complete operation of the cochlea is still not understood, but several mechanisms for sending information to the brain are known to be involved. For sounds below about 1000 Hz, the nerves send signals at the same frequency as the sound. For frequencies greater than about 1000 Hz, the nerves signal frequency by position. There is a structure to the cilia, and there are connections between nerve cells that perform signal processing before information is sent to the brain. Intensity information is partly indicated by the number of nerve signals and by volleys of signals. The brain processes the cochlear nerve signals to provide additional information such as source direction (based on time and intensity comparisons of sounds from both ears). Higher-level processing produces many nuances, such as music appreciation.



The inner ear, or cochlea, is a coiled tube about 3 mm in diameter and 3 cm in length if uncoiled. When the oval window is forced inward, as shown, a pressure wave travels through the perilymph in the direction of the arrows, stimulating nerves at the base of cilia in the organ of Corti.

Hearing losses can occur because of problems in the middle or inner ear. Conductive losses in the middle ear can be partially overcome by sending sound vibrations to the cochlea through the skull. Hearing aids for this purpose usually press against the bone behind the ear, rather than simply amplifying the sound sent into the ear canal as many hearing aids do. Damage to the nerves in the cochlea is not repairable, but amplification can partially compensate. There is a risk that amplification will produce further damage. Another common failure in the cochlea is damage or loss of the cilia but with nerves remaining functional. Cochlear implants that stimulate the nerves directly are now available and widely accepted. Over 100,000 implants are in use, in about equal numbers of adults and children.

The cochlear implant was pioneered in Melbourne, Australia, by Graeme Clark in the 1970s for his deaf father. The implant consists of three external components and two internal components. The external components are a microphone for picking up sound and converting it into an electrical signal, a speech processor to select certain frequencies and a transmitter to transfer the signal to the internal components through electromagnetic induction. The internal components consist of a receiver/transmitter secured in the bone beneath the skin, which converts the signals into electric impulses and sends them through an internal cable to the cochlea and an array of about 24 electrodes wound through the cochlea. These electrodes in turn send the impulses directly into the brain. The electrodes basically emulate the cilia.

Exercise:

Check Your Understanding

Problem:

Are ultrasound and infrasound imperceptible to all hearing organisms? Explain your answer.

Solution:

No, the range of perceptible sound is based in the range of human hearing. Many other organisms perceive either infrasound or ultrasound.

Section Summary

- The range of audible frequencies is 20 to 20,000 Hz.
- Those sounds above 20,000 Hz are ultrasound, whereas those below 20 Hz are infrasound.
- The perception of frequency is pitch.
- The perception of intensity is loudness.
- Loudness has units of phons.

Conceptual Questions

Exercise:

Problem:

Why can a hearing test show that your threshold of hearing is 0 dB at 250 Hz, when [link] implies that no one can hear such a frequency at less than 20 dB?

Problems & Exercises

Exercise:

Problem:

The factor of 10^{-12} in the range of intensities to which the ear can respond, from threshold to that causing damage after brief exposure, is truly remarkable. If you could measure distances over the same range with a single instrument and the smallest distance you could measure was 1 mm, what would the largest be?

Solution:
Equation:

$$1 \times 10^{6}\ \text{km}$$

Exercise:

Problem:

The frequencies to which the ear responds vary by a factor of 10^3. Suppose the speedometer on your car measured speeds differing by the same factor of 10^3, and the greatest speed it reads is 90.0 mi/h. What would be the slowest nonzero speed it could read?

Exercise:

Problem:

What are the closest frequencies to 500 Hz that an average person can clearly distinguish as being different in frequency from 500 Hz? The sounds are not present simultaneously.

Solution:

498.5 or 501.5 Hz

Exercise:

Problem:

Can the average person tell that a 2002-Hz sound has a different frequency than a 1999-Hz sound without playing them simultaneously?

Exercise:

Problem:

If your radio is producing an average sound intensity level of 85 dB, what is the next lowest sound intensity level that is clearly less intense?

Solution:

82 dB

Exercise:

Problem:

Can you tell that your roommate turned up the sound on the TV if its average sound intensity level goes from 70 to 73 dB?

Exercise:

Problem:

Based on the graph in [link], what is the threshold of hearing in decibels for frequencies of 60, 400, 1000, 4000, and 15,000 Hz? Note that many AC electrical appliances produce 60 Hz, music is commonly 400 Hz, a reference frequency is 1000 Hz, your maximum sensitivity is near 4000 Hz, and many older TVs produce a 15,750 Hz whine.

Solution:

approximately 48, 9, 0, –7, and 20 dB, respectively

Exercise:

Problem:

What sound intensity levels must sounds of frequencies 60, 3000, and 8000 Hz have in order to have the same loudness as a 40-dB sound of frequency 1000 Hz (that is, to have a loudness of 40 phons)?

Exercise:

Problem:

What is the approximate sound intensity level in decibels of a 600-Hz tone if it has a loudness of 20 phons? If it has a loudness of 70 phons?

Solution:

(a) 23 dB

(b) 70 dB

Exercise:

Problem:

(a) What are the loudnesses in phons of sounds having frequencies of 200, 1000, 5000, and 10,000 Hz, if they are all at the same 60.0-dB sound intensity level? (b) If they are all at 110 dB? (c) If they are all at 20.0 dB?

Exercise:

Problem:

Suppose a person has a 50-dB hearing loss at all frequencies. By how many factors of 10 will low-intensity sounds need to be amplified to seem normal to this person? Note that smaller amplification is appropriate for more intense sounds to avoid further hearing damage.

Solution:

Five factors of 10

Exercise:

Problem:

If a woman needs an amplification of 5.0×10^{12} times the threshold intensity to enable her to hear at all frequencies, what is her overall hearing loss in dB? Note that smaller amplification is appropriate for more intense sounds to avoid further damage to her hearing from levels above 90 dB.

Exercise:

Problem:

(a) What is the intensity in watts per meter squared of a just barely audible 200-Hz sound? (b) What is the intensity in watts per meter squared of a barely audible 4000-Hz sound?

Solution:

(a) $2 \times 10^{-10}\ \mathrm{W/m^2}$

(b) $2 \times 10^{-13}\ \mathrm{W/m^2}$

Exercise:

Problem:

(a) Find the intensity in watts per meter squared of a 60.0-Hz sound having a loudness of 60 phons. (b) Find the intensity in watts per meter squared of a 10,000-Hz sound having a loudness of 60 phons.

Exercise:

Problem:

A person has a hearing threshold 10 dB above normal at 100 Hz and 50 dB above normal at 4000 Hz. How much more intense must a 100-Hz tone be than a 4000-Hz tone if they are both barely audible to this person?

Solution:

2.5

Exercise:

Problem:

A child has a hearing loss of 60 dB near 5000 Hz, due to noise exposure, and normal hearing elsewhere. How much more intense is a 5000-Hz tone than a 400-Hz tone if they are both barely audible to the child?

Exercise:

Problem:

What is the ratio of intensities of two sounds of identical frequency if the first is just barely discernible as louder to a person than the second?

Solution:

1.26

Glossary

loudness
: the perception of sound intensity

timbre
: number and relative intensity of multiple sound frequencies

note
: basic unit of music with specific names, combined to generate tunes

tone
: number and relative intensity of multiple sound frequencies

phon
: the numerical unit of loudness

ultrasound
: sounds above 20,000 Hz

infrasound
: sounds below 20 Hz

Ultrasound

- Define acoustic impedance and intensity reflection coefficient.
- Describe medical and other uses of ultrasound technology.
- Calculate acoustic impedance using density values and the speed of ultrasound.
- Calculate the velocity of a moving object using Doppler-shifted ultrasound.

Ultrasound is used in medicine to painlessly and noninvasively monitor patient health and diagnose a wide range of disorders. (credit: abbybatchelder, Flickr)

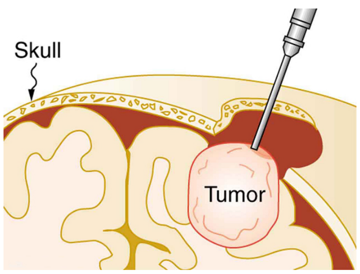
Any sound with a frequency above 20,000 Hz (or 20 kHz)—that is, above the highest audible frequency—is defined to be ultrasound. In practice, it is possible to create ultrasound frequencies up to more than a gigahertz. (Higher frequencies are difficult to create; furthermore, they propagate poorly because they are very strongly absorbed.) Ultrasound has a tremendous number of applications, which range from burglar alarms to use in cleaning delicate objects to the guidance systems of bats. We begin our discussion of ultrasound with some of its applications in medicine, in which it is used extensively both for diagnosis and for therapy.

Note:
Characteristics of Ultrasound
The characteristics of ultrasound, such as frequency and intensity, are wave properties common to all types of waves. Ultrasound also has a wavelength that limits the fineness of detail it can detect. This characteristic is true of all waves. We can never observe details significantly smaller than the wavelength of our probe; for example,

we will never see individual atoms with visible light, because the atoms are so small compared with the wavelength of light.

Ultrasound in Medical Therapy

Ultrasound, like any wave, carries energy that can be absorbed by the medium carrying it, producing effects that vary with intensity. When focused to intensities of 10^3 to $10^5\ \mathrm{W/m^2}$, ultrasound can be used to shatter gallstones or pulverize cancerous tissue in surgical procedures. (See [link].) Intensities this great can damage individual cells, variously causing their protoplasm to stream inside them, altering their permeability, or rupturing their walls through *cavitation*. Cavitation is the creation of vapor cavities in a fluid—the longitudinal vibrations in ultrasound alternatively compress and expand the medium, and at sufficient amplitudes the expansion separates molecules. Most cavitation damage is done when the cavities collapse, producing even greater shock pressures.



The tip of this small probe oscillates at 23 kHz with such a large amplitude that it pulverizes tissue on contact. The debris is then aspirated. The speed of the tip may exceed the speed of sound in tissue, thus creating shock waves and cavitation, rather than a smooth

simple harmonic oscillator–type wave.

Most of the energy carried by high-intensity ultrasound in tissue is converted to thermal energy. In fact, intensities of 10^3 to $10^4\ \mathrm{W/m^2}$ are commonly used for deep-heat treatments called ultrasound diathermy. Frequencies of 0.8 to 1 MHz are typical. In both athletics and physical therapy, ultrasound diathermy is most often applied to injured or overworked muscles to relieve pain and improve flexibility. Skill is needed by the therapist to avoid "bone burns" and other tissue damage caused by overheating and cavitation, sometimes made worse by reflection and focusing of the ultrasound by joint and bone tissue.

In some instances, you may encounter a different decibel scale, called the sound *pressure* level, when ultrasound travels in water or in human and other biological tissues. We shall not use the scale here, but it is notable that numbers for sound pressure levels range 60 to 70 dB higher than you would quote for β, the sound intensity level used in this text. Should you encounter a sound pressure level of 220 decibels, then, it is not an astronomically high intensity, but equivalent to about 155 dB—high enough to destroy tissue, but not as unreasonably high as it might seem at first.

Ultrasound in Medical Diagnostics

When used for imaging, ultrasonic waves are emitted from a transducer, a crystal exhibiting the piezoelectric effect (the expansion and contraction of a substance when a voltage is applied across it, causing a vibration of the crystal). These high-frequency vibrations are transmitted into any tissue in contact with the transducer. Similarly, if a pressure is applied to the crystal (in the form of a wave reflected off tissue layers), a voltage is produced which can be recorded. The crystal therefore acts as both a transmitter and a receiver of sound. Ultrasound is also partially absorbed by tissue on its path, both on its journey away from the transducer and on its return journey. From the time between when the original signal is sent and when the reflections from various boundaries between media are received, (as well as a measure of the intensity loss of the signal), the nature and position of each boundary between tissues and organs may be deduced.

Reflections at boundaries between two different media occur because of differences in a characteristic known as the **acoustic impedance** Z of each substance. Impedance is defined as

Equation:

$$Z = \rho v,$$

where ρ is the density of the medium (in kg/m^3) and v is the speed of sound through the medium (in m/s). The units for Z are therefore $\text{kg/(m}^2 \cdot \text{s)}$.

[link] shows the density and speed of sound through various media (including various soft tissues) and the associated acoustic impedances. Note that the acoustic impedances for soft tissue do not vary much but that there is a big difference between the acoustic impedance of soft tissue and air and also between soft tissue and bone.

Medium	**Density (kg/m^3)**	**Speed of Ultrasound (m/s)**	**Acoustic Impedance** $\left(\text{kg/}\left(\text{m}^2 \cdot \text{s}\right)\right)$
Air	1.3	330	429
Water	1000	1500	1.5×10^6
Blood	1060	1570	1.66×10^6
Fat	925	1450	1.34×10^6
Muscle (average)	1075	1590	1.70×10^6
Bone (varies)	1400–1900	4080	5.7×10^6 to 7.8×10^6
Barium titanate (transducer material)	5600	5500	30.8×10^6

The Ultrasound Properties of Various Media, Including Soft Tissue Found in the Body

At the boundary between media of different acoustic impedances, some of the wave energy is reflected and some is transmitted. The greater the *difference* in acoustic impedance between the two media, the greater the reflection and the smaller the transmission.

The **intensity reflection coefficient** a is defined as the ratio of the intensity of the reflected wave relative to the incident (transmitted) wave. This statement can be written mathematically as

Equation:

$$a = \frac{\left(Z_2 - Z_1\right)^2}{\left(Z_1 + Z_2\right)^2},$$

where Z_1 and Z_2 are the acoustic impedances of the two media making up the boundary. A reflection coefficient of zero (corresponding to total transmission and no reflection) occurs when the acoustic impedances of the two media are the same. An impedance "match" (no reflection) provides an efficient coupling of sound energy from one medium to another. The image formed in an ultrasound is made by tracking reflections (as shown in [link]) and mapping the intensity of the reflected sound waves in a two-dimensional plane.

Example:

Calculate Acoustic Impedance and Intensity Reflection Coefficient: Ultrasound and Fat Tissue

(a) Using the values for density and the speed of ultrasound given in [link], show that the acoustic impedance of fat tissue is indeed $1.34 \times 10^6\ \text{kg/(m}^2\text{·s)}$.

(b) Calculate the intensity reflection coefficient of ultrasound when going from fat to muscle tissue.

Strategy for (a)

The acoustic impedance can be calculated using $Z = \rho\text{v}$ and the values for ρ and v found in [link].

Solution for (a)

(1) Substitute known values from [link] into $Z = \rho\text{v}$.

Equation:

$$Z = \rho v = \left(925\ \text{kg/m}^3\right)\left(1450\ \text{m/s}\right)$$

(2) Calculate to find the acoustic impedance of fat tissue.

Equation:

$$1.34 \times 10^6 \text{ kg/(m}^2\cdot\text{s)}$$

This value is the same as the value given for the acoustic impedance of fat tissue.

Strategy for (b)

The intensity reflection coefficient for any boundary between two media is given by $a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}$, and the acoustic impedance of muscle is given in [link].

Solution for (b)

Substitute known values into $a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}$ to find the intensity reflection coefficient:

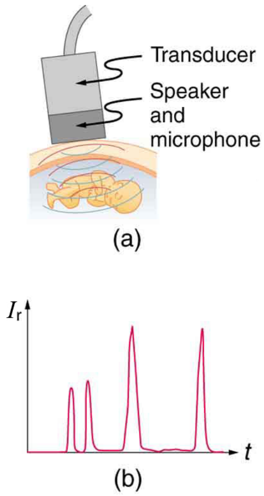
Equation:

$$a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2} = \frac{\left(1.34 \times 10^6 \text{ kg/(m}^2\cdot\text{ s)} - 1.70 \times 10^6 \text{ kg/(m}^2\cdot\text{ s)}\right)^2}{\left(1.70 \times 10^6 \text{ kg/(m}^2\cdot\text{ s)} + 1.34 \times 10^6 \text{ kg/(m}^2\cdot\text{ s)}\right)^2} = 0.014$$

Discussion

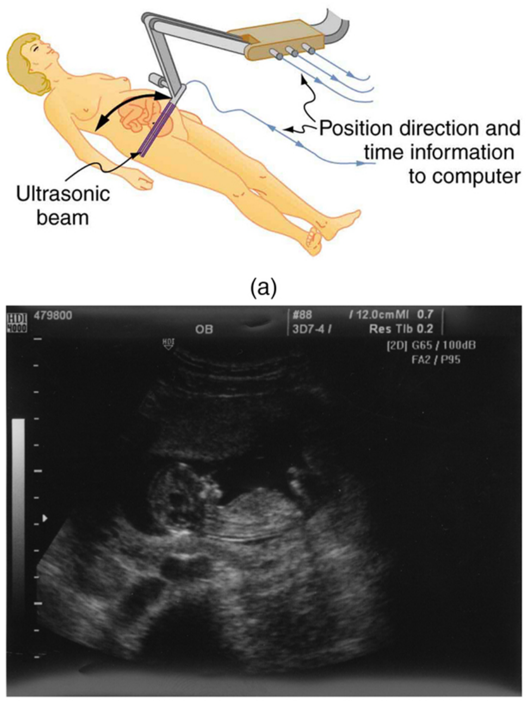
This result means that only 1.4% of the incident intensity is reflected, with the remaining being transmitted.

The applications of ultrasound in medical diagnostics have produced untold benefits with no known risks. Diagnostic intensities are too low (about 10^{-2} W/m^2) to cause thermal damage. More significantly, ultrasound has been in use for several decades and detailed follow-up studies do not show evidence of ill effects, quite unlike the case for x-rays.



(a) An ultrasound speaker doubles as a microphone. Brief bleeps are broadcast, and echoes are recorded from various depths. (b) Graph of echo intensity versus time. The time for echoes to return is directly proportional to the distance of the reflector, yielding this information noninvasively.

The most common ultrasound applications produce an image like that shown in [link]. The speaker-microphone broadcasts a directional beam, sweeping the beam across the area of interest. This is accomplished by having multiple ultrasound sources in the probe’s head, which are phased to interfere constructively in a given, adjustable direction. Echoes are measured as a function of position as well as depth. A computer constructs an image that reveals the shape and density of internal structures.



(a) An ultrasonic image is produced by sweeping the ultrasonic beam across the area of interest, in this case the woman’s abdomen. Data are recorded and analyzed in a computer, providing a two-dimensional image. (b) Ultrasound image of 12-week-old fetus. (credit: Margaret W. Carruthers, Flickr)

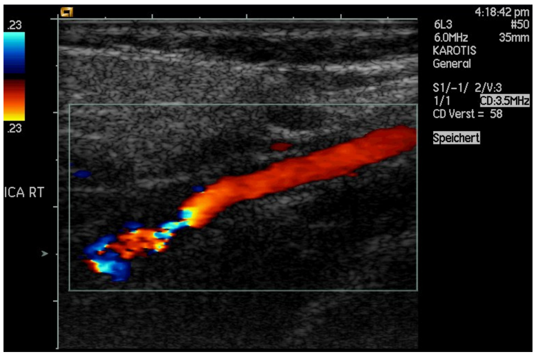
How much detail can ultrasound reveal? The image in [link] is typical of low-cost systems, but that in [link] shows the remarkable detail possible with more advanced systems, including 3D imaging. Ultrasound today is commonly used in prenatal care. Such imaging can be used to see if the fetus is developing at a normal rate, and help in the determination of serious problems early in the pregnancy. Ultrasound is also in wide use to image the chambers of the heart and the flow of blood within the beating heart, using the Doppler effect (echocardiology).

Whenever a wave is used as a probe, it is very difficult to detect details smaller than its wavelength λ. Indeed, current technology cannot do quite this well. Abdominal scans may use a 7-MHz frequency, and the speed of sound in tissue is about 1540 m/s—so the wavelength limit to detail would be $\lambda = \frac{v_{\text{w}}}{f} = \frac{1540\text{ m/s}}{7\times10^6\text{ Hz}} = 0.22\text{ mm}$. In practice, 1-mm detail is attainable, which is sufficient for many purposes. Higher-frequency ultrasound would allow greater detail, but it does not penetrate as well as lower frequencies do. The accepted rule of thumb is that you can effectively scan to a depth of about 500λ into tissue. For 7 MHz, this penetration limit is $500 \times 0.22\text{ mm}$, which is 0.11 m. Higher frequencies may be employed in smaller organs, such as the eye, but are not practical for looking deep into the body.

A 3D ultrasound image of a fetus. As well as for the detection of any abnormalities, such scans have also been shown to be useful for strengthening the emotional bonding between parents and their unborn child. (credit: Jennie Cu, Wikimedia Commons)

In addition to shape information, ultrasonic scans can produce density information superior to that found in X-rays, because the intensity of a reflected sound is related to changes in density. Sound is most strongly reflected at places where density changes are greatest.

Another major use of ultrasound in medical diagnostics is to detect motion and determine velocity through the Doppler shift of an echo, known as **Doppler-shifted ultrasound**. This technique is used to monitor fetal heartbeat, measure blood velocity, and detect occlusions in blood vessels, for example. (See [link].) The magnitude of the Doppler shift in an echo is directly proportional to the velocity of whatever reflects the sound. Because an echo is involved, there is actually a double shift. The first occurs because the reflector (say a fetal heart) is a moving observer and receives a Doppler-shifted frequency. The reflector then acts as a moving source, producing a second Doppler shift.



This Doppler-shifted ultrasonic image of a partially occluded artery uses color to indicate velocity. The highest velocities are in red, while the lowest are blue. The blood must move faster through the constriction to carry the same flow. (credit: Arning C, Grzyska U, Wikimedia Commons)

A clever technique is used to measure the Doppler shift in an echo. The frequency of the echoed sound is superimposed on the broadcast frequency, producing beats. The beat frequency is $F_{\text{B}} = \mid f_1 - f_2 \mid$, and so it is directly proportional to the Doppler shift $(f_1 - f_2)$ and hence, the reflector's velocity. The advantage in this technique is that the Doppler shift is small (because the reflector's velocity is small), so that great accuracy would be needed to measure the shift directly. But measuring the beat frequency is easy, and it is not affected if the broadcast frequency varies somewhat. Furthermore, the beat frequency is in the audible range and can be amplified for audio feedback to the medical observer.

Note:

Uses for Doppler-Shifted Radar

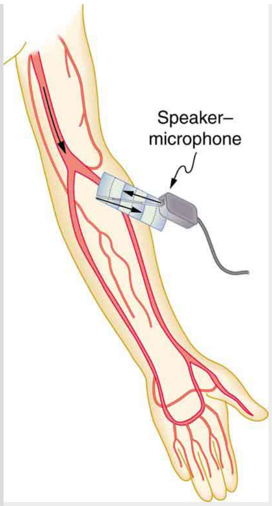
Doppler-shifted radar echoes are used to measure wind velocities in storms as well as aircraft and automobile speeds. The principle is the same as for Doppler-shifted ultrasound. There is evidence that bats and dolphins may also sense the velocity of an object (such as prey) reflecting their ultrasound signals by observing its Doppler shift.

Example:

Calculate Velocity of Blood: Doppler-Shifted Ultrasound

Ultrasound that has a frequency of 2.50 MHz is sent toward blood in an artery that is moving toward the source at 20.0 cm/s, as illustrated in [link]. Use the speed of sound in human tissue as 1540 m/s. (Assume that the frequency of 2.50 MHz is accurate to seven significant figures.)

a. What frequency does the blood receive?
b. What frequency returns to the source?
c. What beat frequency is produced if the source and returning frequencies are mixed?



Ultrasound is partly reflected by blood cells and plasma back toward the speaker-microphone. Because the cells are moving, two Doppler shifts are produced—one for blood as a moving observer, and the other for the reflected sound coming from a moving source. The magnitude of the shift is directly proportional to blood velocity.

Strategy

The first two questions can be answered using $f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right)$ and $f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right)$ for the Doppler shift. The last question asks for beat frequency, which is the difference between the original and returning frequencies.

Solution for (a)

(1) Identify knowns:

- The blood is a moving observer, and so the frequency it receives is given by
  Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right).$$

- v_{b} is the blood velocity (v_{obs} here) and the plus sign is chosen because the motion is toward the source.

(2) Enter the given values into the equation.

Equation:

$$f_{\text{obs}} = (2{,}500{,}000\ \text{Hz})\left(\frac{1540\ \text{m/s} + 0.2\ \text{m/s}}{1540\ \text{m/s}}\right)$$

(3) Calculate to find the frequency: 2,500,325 Hz.

Solution for (b)

(1) Identify knowns:

- The blood acts as a moving source.
- The microphone acts as a stationary observer.
- The frequency leaving the blood is 2,500,325 Hz, but it is shifted upward as given by
  Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} - v_{\text{b}}}\right).$$

  f_{obs} is the frequency received by the speaker-microphone.
- The source velocity is v_{b}.
- The minus sign is used because the motion is toward the observer.

The minus sign is used because the motion is toward the observer.
(2) Enter the given values into the equation:
Equation:

$$f_{\text{obs}} = (2{,}500{,}325\text{ Hz})\left(\frac{1540\text{ m/s}}{1540\text{ m/s} - 0.200\text{ m/s}}\right)$$

(3) Calculate to find the frequency returning to the source: 2,500,649 Hz.
Solution for (c)
(1) Identify knowns:

- The beat frequency is simply the absolute value of the difference between f_{s} and f_{obs}, as stated in:
  Equation:

  $$f_{\text{B}} = \mid f_{\text{obs}} - f_{\text{s}} \mid .$$

(2) Substitute known values:
Equation:

$$\mid 2{,}500{,}649\text{ Hz} - 2{,}500{,}000\text{ Hz} \mid$$

(3) Calculate to find the beat frequency: 649 Hz.
Discussion
The Doppler shifts are quite small compared with the original frequency of 2.50 MHz. It is far easier to measure the beat frequency than it is to measure the echo frequency with an accuracy great enough to see shifts of a few hundred hertz out of a couple of megahertz. Furthermore, variations in the source frequency do not greatly affect the beat frequency, because both f_{s} and f_{obs}would increase or decrease. Those changes subtract out in $f_{\text{B}} = \mid f_{\text{obs}} - f_{\text{s}} \mid$.

Note:
Industrial and Other Applications of Ultrasound
Industrial, retail, and research applications of ultrasound are common. A few are discussed here. Ultrasonic cleaners have many uses. Jewelry, machined parts, and other objects that have odd shapes and crevices are immersed in a cleaning fluid that is agitated with ultrasound typically about 40 kHz in frequency. The intensity is great enough to cause cavitation, which is responsible for most of the cleansing action. Because cavitation-produced shock pressures are large and well transmitted in a fluid,

they reach into small crevices where even a low-surface-tension cleaning fluid might not penetrate.

Sonar is a familiar application of ultrasound. Sonar typically employs ultrasonic frequencies in the range from 30.0 to 100 kHz. Bats, dolphins, submarines, and even some birds use ultrasonic sonar. Echoes are analyzed to give distance and size information both for guidance and finding prey. In most sonar applications, the sound reflects quite well because the objects of interest have significantly different density than the medium in which they travel. When the Doppler shift is observed, velocity information can also be obtained. Submarine sonar can be used to obtain such information, and there is evidence that some bats also sense velocity from their echoes.

Similarly, there are a range of relatively inexpensive devices that measure distance by timing ultrasonic echoes. Many cameras, for example, use such information to focus automatically. Some doors open when their ultrasonic ranging devices detect a nearby object, and certain home security lights turn on when their ultrasonic rangers observe motion. Ultrasonic "measuring tapes" also exist to measure such things as room dimensions. Sinks in public restrooms are sometimes automated with ultrasound devices to turn faucets on and off when people wash their hands. These devices reduce the spread of germs and can conserve water.

Ultrasound is used for nondestructive testing in industry and by the military. Because ultrasound reflects well from any large change in density, it can reveal cracks and voids in solids, such as aircraft wings, that are too small to be seen with x-rays. For similar reasons, ultrasound is also good for measuring the thickness of coatings, particularly where there are several layers involved.

Basic research in solid state physics employs ultrasound. Its attenuation is related to a number of physical characteristics, making it a useful probe. Among these characteristics are structural changes such as those found in liquid crystals, the transition of a material to a superconducting phase, as well as density and other properties.

These examples of the uses of ultrasound are meant to whet the appetites of the curious, as well as to illustrate the underlying physics of ultrasound. There are many more applications, as you can easily discover for yourself.

Exercise:
Check Your Understanding

Problem:

Why is it possible to use ultrasound both to observe a fetus in the womb and also to destroy cancerous tumors in the body?

Solution:

Ultrasound can be used medically at different intensities. Lower intensities do not cause damage and are used for medical imaging. Higher intensities can pulverize and destroy targeted substances in the body, such as tumors.

Section Summary

- The acoustic impedance is defined as:
  Equation:
  $$Z = \rho v,$$
  ρ is the density of a medium through which the sound travels and v is the speed of sound through that medium.
- The intensity reflection coefficient a, a measure of the ratio of the intensity of the wave reflected off a boundary between two media relative to the intensity of the incident wave, is given by
  Equation:
  $$a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}.$$
- The intensity reflection coefficient is a unitless quantity.

Conceptual Questions

Exercise:

Problem:

If audible sound follows a rule of thumb similar to that for ultrasound, in terms of its absorption, would you expect the high or low frequencies from your neighbor’s stereo to penetrate into your house? How does this expectation compare with your experience?

Exercise:

Problem:

Elephants and whales are known to use infrasound to communicate over very large distances. What are the advantages of infrasound for long distance communication?

Exercise:

Problem:

It is more difficult to obtain a high-resolution ultrasound image in the abdominal region of someone who is overweight than for someone who has a slight build. Explain why this statement is accurate.

Exercise:

Problem:

Suppose you read that 210-dB ultrasound is being used to pulverize cancerous tumors. You calculate the intensity in watts per centimeter squared and find it is unreasonably high ($10^5\ \text{W/cm}^2$). What is a possible explanation?

Problems & Exercises

Unless otherwise indicated, for problems in this section, assume that the speed of sound through human tissues is 1540 m/s.

Exercise:

Problem:

What is the sound intensity level in decibels of ultrasound of intensity $10^5\ \text{W/m}^2$, used to pulverize tissue during surgery?

Solution:

170 dB

Exercise:

Problem:

Is 155-dB ultrasound in the range of intensities used for deep heating? Calculate the intensity of this ultrasound and compare this intensity with values quoted in the text.

Exercise:

Problem:

Find the sound intensity level in decibels of $2.00 \times 10^{-2}\ \text{W/m}^2$ ultrasound used in medical diagnostics.

Solution:

103 dB

Exercise:

Problem:

The time delay between transmission and the arrival of the reflected wave of a signal using ultrasound traveling through a piece of fat tissue was 0.13 ms. At what depth did this reflection occur?

Exercise:

Problem:

In the clinical use of ultrasound, transducers are always coupled to the skin by a thin layer of gel or oil, replacing the air that would otherwise exist between the transducer and the skin. (a) Using the values of acoustic impedance given in [link] calculate the intensity reflection coefficient between transducer material and air. (b) Calculate the intensity reflection coefficient between transducer material and gel (assuming for this problem that its acoustic impedance is identical to that of water). (c) Based on the results of your calculations, explain why the gel is used.

Solution:

(a) 1.00

(b) 0.823

(c) Gel is used to facilitate the transmission of the ultrasound between the transducer and the patient’s body.

Exercise:

Problem:

(a) Calculate the minimum frequency of ultrasound that will allow you to see details as small as 0.250 mm in human tissue. (b) What is the effective depth to which this sound is effective as a diagnostic probe?

Exercise:

Problem:

(a) Find the size of the smallest detail observable in human tissue with 20.0-MHz ultrasound. (b) Is its effective penetration depth great enough to examine the entire eye (about 3.00 cm is needed)? (c) What is the wavelength of such ultrasound in 0°C air?

Solution:

(a) $77.0\ \mu\text{m}$

(b) Effective penetration depth = 3.85 cm, which is enough to examine the eye.

(c) $16.6\ \mu\text{m}$

Exercise:

Problem:

(a) Echo times are measured by diagnostic ultrasound scanners to determine distances to reflecting surfaces in a patient. What is the difference in echo times for tissues that are 3.50 and 3.60 cm beneath the surface? (This difference is the minimum resolving time for the scanner to see details as small as 0.100 cm, or 1.00 mm. Discrimination of smaller time differences is needed to see smaller details.) (b) Discuss whether the period T of this ultrasound must be smaller than the minimum time resolution. If so, what is the minimum frequency of the ultrasound and is that out of the normal range for diagnostic ultrasound?

Exercise:

Problem:

(a) How far apart are two layers of tissue that produce echoes having round-trip times (used to measure distances) that differ by $0.750\ \mu\text{s}$? (b) What minimum frequency must the ultrasound have to see detail this small?

Solution:

(a) $5.78 \times 10^{-4}\ \text{m}$

(b) $2.67 \times 10^{6}\ \text{Hz}$

Exercise:

Problem:

(a) A bat uses ultrasound to find its way among trees. If this bat can detect echoes 1.00 ms apart, what minimum distance between objects can it detect? (b) Could this distance explain the difficulty that bats have finding an open door when they accidentally get into a house?

Exercise:

Problem:

A dolphin is able to tell in the dark that the ultrasound echoes received from two sharks come from two different objects only if the sharks are separated by 3.50 m, one being that much farther away than the other. (a) If the ultrasound has a frequency of 100 kHz, show this ability is not limited by its wavelength. (b) If this ability is due to the dolphin’s ability to detect the arrival times of echoes, what is the minimum time difference the dolphin can perceive?

Solution:

(a) $v_{\text{w}} = 1540\ \text{m/s} = f\lambda \Rightarrow \lambda = \frac{1540\ \text{m/s}}{100\times10^{3}\ \text{Hz}} = 0.0154\ \text{m} < 3.50\ \text{m}$. Because the wavelength is much shorter than the distance in question, the wavelength is not the limiting factor.

(b) 4.55 ms

Exercise:

Problem:

A diagnostic ultrasound echo is reflected from moving blood and returns with a frequency 500 Hz higher than its original 2.00 MHz. What is the velocity of the blood? (Assume that the frequency of 2.00 MHz is accurate to seven significant figures and 500 Hz is accurate to three significant figures.)

Exercise:

Problem:

Ultrasound reflected from an oncoming bloodstream that is moving at 30.0 cm/s is mixed with the original frequency of 2.50 MHz to produce beats. What is the beat frequency? (Assume that the frequency of 2.50 MHz is accurate to seven significant figures.)

Solution:

974 Hz

(Note: extra digits were retained in order to show the difference.)

Glossary

acoustic impedance
: property of medium that makes the propagation of sound waves more difficult

intensity reflection coefficient
: a measure of the ratio of the intensity of the wave reflected off a boundary between two media relative to the intensity of the incident wave

Doppler-shifted ultrasound
: a medical technique to detect motion and determine velocity through the Doppler shift of an echo

Pressures in the Body

- Explain the concept of pressure the in human body.
- Explain systolic and diastolic blood pressures.
- Describe pressures in the eye, lungs, spinal column, bladder, and skeletal system.

Pressure in the Body

Next to taking a person's temperature and weight, measuring blood pressure is the most common of all medical examinations. Control of high blood pressure is largely responsible for the significant decreases in heart attack and stroke fatalities achieved in the last three decades. The pressures in various parts of the body can be measured and often provide valuable medical indicators. In this section, we consider a few examples together with some of the physics that accompanies them.

[link] lists some of the measured pressures in mm Hg, the units most commonly quoted.

Body system	Gauge pressure in mm Hg
Blood pressures in large arteries (resting)	
Maximum (systolic)	100–140
Minimum (diastolic)	60–90
Blood pressure in large veins	4–15
Eye	12–24
Brain and spinal fluid (lying down)	5–12
Bladder	
While filling	0–25
When full	100–150
Chest cavity between lungs and ribs	−8 to −4
Inside lungs	−2 to +3
Digestive tract	

Body system	Gauge pressure in mm Hg
Esophagus	−2
Stomach	0–20
Intestines	10–20
Middle ear	<1

Typical Pressures in Humans

Blood Pressure

Common arterial blood pressure measurements typically produce values of 120 mm Hg and 80 mm Hg, respectively, for systolic and diastolic pressures. Both pressures have health implications. When systolic pressure is chronically high, the risk of stroke and heart attack is increased. If, however, it is too low, fainting is a problem. **Systolic pressure** increases dramatically during exercise to increase blood flow and returns to normal afterward. This change produces no ill effects and, in fact, may be beneficial to the tone of the circulatory system. **Diastolic pressure** can be an indicator of fluid balance. When low, it may indicate that a person is hemorrhaging internally and needs a transfusion. Conversely, high diastolic pressure indicates a ballooning of the blood vessels, which may be due to the transfusion of too much fluid into the circulatory system. High diastolic pressure is also an indication that blood vessels are not dilating properly to pass blood through. This can seriously strain the heart in its attempt to pump blood.

Blood leaves the heart at about 120 mm Hg but its pressure continues to decrease (to almost 0) as it goes from the aorta to smaller arteries to small veins (see [link]). The pressure differences in the circulation system are caused by blood flow through the system as well as the position of the person. For a person standing up, the pressure in the feet will be larger than at the heart due to the weight of the blood $(P = h\rho g)$. If we assume that the distance between the heart and the feet of a person in an upright position is 1.4 m, then the increase in pressure in the feet relative to that in the heart (for a static column of blood) is given by

Equation:

$$\Delta P = \Delta h\rho g = (1.4\text{ m})\left(1050\text{ kg/m}^3\right)\left(9.80\text{ m/s}^2\right) = 1.4 \times 10^4\text{ Pa} = 108\text{ mm Hg.}$$

Note:

Increase in Pressure in the Feet of a Person

Equation:

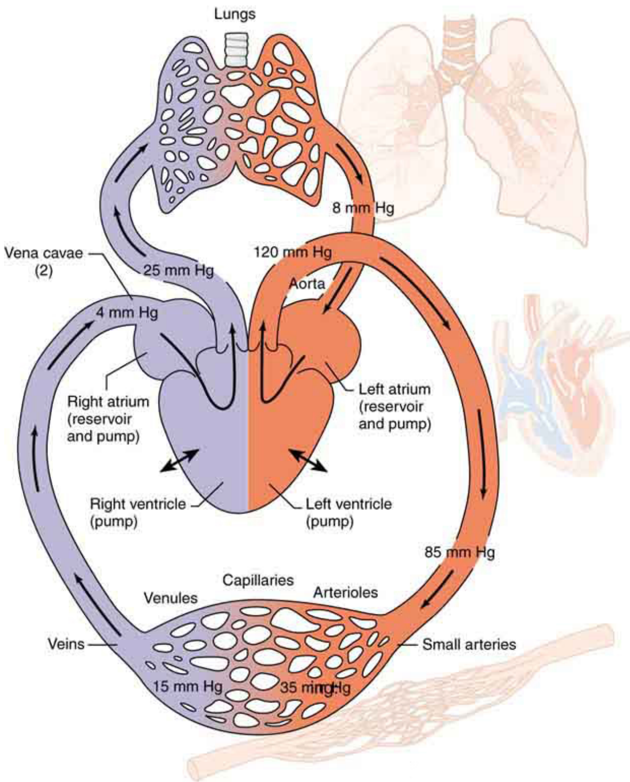
$$\Delta P = \Delta h\rho g = (1.4\text{ m})\left(1050\text{ kg/m}^3\right)\left(9.80\text{ m/s}^2\right) = 1.4 \times 10^4\text{ Pa} = 108\text{ mm Hg.}$$

Standing a long time can lead to an accumulation of blood in the legs and swelling. This is the reason why soldiers who are required to stand still for long periods of time have been known to faint. Elastic bandages around the calf can help prevent this accumulation and can also help provide increased pressure to enable the veins to send blood back up to the heart. For similar reasons, doctors recommend tight stockings for long-haul flights.

Blood pressure may also be measured in the major veins, the heart chambers, arteries to the brain, and the lungs. But these pressures are usually only monitored during surgery or for patients in intensive care since the measurements are invasive. To obtain these pressure measurements, qualified health care workers thread thin tubes, called catheters, into appropriate locations to transmit pressures to external measuring devices.

The heart consists of two pumps—the right side forcing blood through the lungs and the left causing blood to flow through the rest of the body ([link]). Right-heart failure, for example, results in a rise in the pressure in the vena cavae and a drop in pressure in the arteries to the lungs. Left-heart failure results in a rise in the pressure entering the left side of the heart and a drop in aortal pressure. Implications of these and other pressures on flow in the circulatory system will be discussed in more detail in Fluid Dynamics and Its Biological and Medical Applications.

Note:
Two Pumps of the Heart
The heart consists of two pumps—the right side forcing blood through the lungs and the left causing blood to flow through the rest of the body.



Schematic of the circulatory system showing typical pressures. The two pumps in the heart increase pressure and that pressure is reduced as the blood flows through the body. Long-term deviations from these pressures have medical implications discussed in some detail in the Fluid Dynamics and Its Biological and Medical Applications. Only aortal or arterial blood pressure can be measured noninvasively.

Pressure in the Eye

The shape of the eye is maintained by fluid pressure, called **intraocular pressure**, which is normally in the range of 12.0 to 24.0 mm Hg. When the circulation of fluid in the eye is blocked, it can lead to a buildup in pressure, a condition called **glaucoma**. The net pressure can become as great as 85.0 mm Hg, an abnormally large pressure that can permanently damage the optic nerve. To get an idea of the force involved, suppose the back of the eye has an area of $6.0\ \text{cm}^2$, and the net pressure is 85.0 mm Hg. Force is given by $F = \text{PA}$. To get F in newtons, we convert the area to m^2 ($1\ \text{m}^2 = 10^4\ \text{cm}^2$). Then we calculate as follows:

Equation:

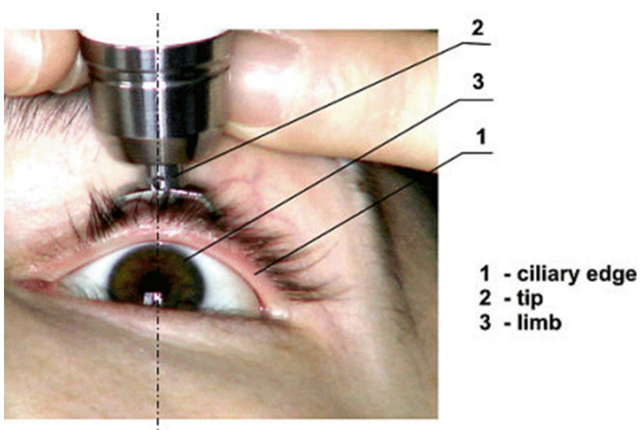
$$F = h\rho gA = \left(85.0 \times 10^{-3}\ \text{m}\right)\left(13.6 \times 10^{3}\ \text{kg/m}^{3}\right)\left(9.80\ \text{m/s}^{2}\right)\left(6.0 \times 10^{-4}\ \text{m}^{2}\right) = 6.8\ \text{N.}$$

Note:
Eye Pressure
The shape of the eye is maintained by fluid pressure, called intraocular pressure. When the circulation of fluid in the eye is blocked, it can lead to a buildup in pressure, a condition called glaucoma. The force is calculated as
Equation:

$$F = h\rho gA = \left(85.0 \times 10^{-3}\ \text{m}\right)\left(13.6 \times 10^{3}\ \text{kg/m}^{3}\right)\left(9.80\ \text{m/s}^{2}\right)\left(6.0 \times 10^{-4}\ \text{m}^{2}\right) = 6.8\ \text{N.}$$

This force is the weight of about a 680-g mass. A mass of 680 g resting on the eye (imagine 1.5 lb resting on your eye) would be sufficient to cause it damage. (A normal force here would be the weight of about 120 g, less than one-quarter of our initial value.)

People over 40 years of age are at greatest risk of developing glaucoma and should have their intraocular pressure tested routinely. Most measurements involve exerting a force on the (anesthetized) eye over some area (a pressure) and observing the eye's response. A noncontact approach uses a puff of air and a measurement is made of the force needed to indent the eye ([link]). If the intraocular pressure is high, the eye will deform less and rebound more vigorously than normal. Excessive intraocular pressures can be detected reliably and sometimes controlled effectively.



The intraocular eye pressure can be read with a tonometer. (credit: DevelopAll at the Wikipedia Project.)

Example:
Calculating Gauge Pressure and Depth: Damage to the Eardrum
Suppose a 3.00-N force can rupture an eardrum. (a) If the eardrum has an area of $1.00\ \mathrm{cm}^2$, calculate the maximum tolerable gauge pressure on the eardrum in newtons per meter squared and convert it to millimeters of mercury. (b) At what depth in freshwater would this person's eardrum rupture, assuming the gauge pressure in the middle ear is zero?
Strategy for (a)
The pressure can be found directly from its definition since we know the force and area. We are looking for the gauge pressure.
Solution for (a)
Equation:

$$P_{\mathrm{g}} = F/A = 3.00\ \mathrm{N}/(1.00 \times 10^{-4}\ \mathrm{m}^2) = 3.00 \times 10^4\ \mathrm{N/m}^2.$$

We now need to convert this to units of mm Hg:
Equation:

$$P_{\mathrm{g}} = 3.0 \times 10^4\ \mathrm{N/m}^2 \left(\frac{1.0\ \mathrm{mm\ Hg}}{133\ \mathrm{N/m}^2} \right) = 226\ \mathrm{mm\ Hg}.$$

Strategy for (b)
Here we will use the fact that the water pressure varies linearly with depth h below the surface.
Solution for (b)
$P = h\rho g$ and therefore $h = P/\rho g$. Using the value above for P, we have
Equation:

$$h = \frac{3.0 \times 10^4\ \mathrm{N/m}^2}{(1.00 \times 10^3\ \mathrm{kg/m}^3)(9.80\ \mathrm{m/s}^2)} = 3.06\ \mathrm{m}.$$

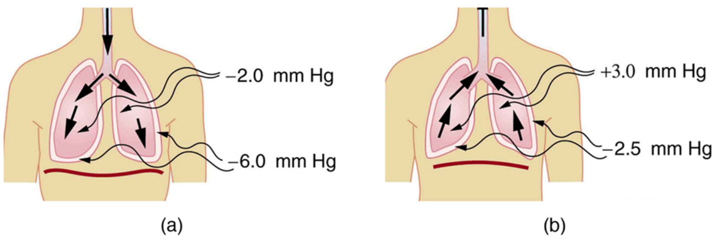
Discussion
Similarly, increased pressure exerted upon the eardrum from the middle ear can arise when an infection causes a fluid buildup.

Pressure Associated with the Lungs

The pressure inside the lungs increases and decreases with each breath. The pressure drops to below atmospheric pressure (negative gauge pressure) when you inhale, causing air to flow into the lungs. It increases above atmospheric pressure (positive gauge pressure) when you exhale, forcing air out.

Lung pressure is controlled by several mechanisms. Muscle action in the diaphragm and rib cage is necessary for inhalation; this muscle action increases the volume of the lungs thereby reducing the pressure within them [link]. Surface tension in the alveoli creates a positive pressure opposing inhalation. (See Cohesion and Adhesion in Liquids: Surface Tension and Capillary Action.) You can exhale without muscle action by letting surface tension in the alveoli create its own positive pressure. Muscle action can add to this positive pressure to produce forced exhalation, such as when you blow up a balloon, blow out a candle, or cough.

The lungs, in fact, would collapse due to the surface tension in the alveoli, if they were not attached to the inside of the chest wall by liquid adhesion. The gauge pressure in the liquid attaching the lungs to the inside of the chest wall is thus negative, ranging from -4 to -8 mm Hg during exhalation and inhalation, respectively. If air is allowed to enter the chest cavity, it breaks the attachment, and one or both lungs may collapse. Suction is applied to the chest cavity of surgery patients and trauma victims to reestablish negative pressure and inflate the lungs.



(a) During inhalation, muscles expand the chest, and the diaphragm moves downward, reducing pressure inside the lungs to less than atmospheric (negative gauge pressure). Pressure between the lungs and chest wall is even lower to overcome the positive pressure created by surface tension in the lungs. (b) During gentle exhalation, the muscles simply relax and surface tension in the alveoli creates a positive pressure inside the lungs, forcing air out. Pressure between the chest wall and lungs remains negative to keep them attached to the chest wall, but it is less negative than during inhalation.

Other Pressures in the Body

Spinal Column and Skull

Normally, there is a 5- to12-mm Hg pressure in the fluid surrounding the brain and filling the spinal column. This cerebrospinal fluid serves many purposes, one of which is to supply flotation to the brain. The buoyant force supplied by the fluid nearly equals the weight of the brain, since their densities are nearly equal. If there is a loss of fluid, the brain rests on the inside of the skull, causing severe headaches, constricted blood flow, and serious damage. Spinal fluid pressure is measured by means of a needle inserted between vertebrae that transmits the pressure to a suitable measuring device.

Bladder Pressure

This bodily pressure is one of which we are often aware. In fact, there is a relationship between our awareness of this pressure and a subsequent increase in it. Bladder pressure climbs steadily from zero to about 25 mm Hg as the bladder fills to its normal capacity of $500\ \mathrm{cm}^3$. This pressure triggers the **micturition reflex**, which stimulates the feeling of needing to urinate. What is more, it also causes muscles around the bladder to contract, raising the pressure to over 100 mm Hg, accentuating the sensation. Coughing, straining, tensing in cold weather, wearing tight clothes, and experiencing simple nervous tension all can increase bladder pressure and trigger this reflex. So can the weight of a pregnant woman’s fetus, especially if it is kicking vigorously or pushing down with its head! Bladder pressure can be measured by a catheter or by inserting a needle through the bladder wall and transmitting the pressure to an appropriate measuring device. One hazard of high bladder pressure (sometimes created by an obstruction), is that such pressure can force urine back into the kidneys, causing potentially severe damage.

Pressures in the Skeletal System

These pressures are the largest in the body, due both to the high values of initial force, and the small areas to which this force is applied, such as in the joints.. For example, when a person lifts an object improperly, a force of 5000 N may be created between vertebrae in the spine, and this may be applied to an area as small as $10\ \mathrm{cm}^2$. The pressure created is $P = F/A = (5000\ \mathrm{N})/(10^{-3}\ \mathrm{m}^2) = 5.0 \times 10^6\ \mathrm{N/m}^2$ or about 50 atm! This pressure can damage both the spinal discs (the cartilage between vertebrae), as well as the bony vertebrae themselves. Even under normal circumstances, forces between vertebrae in the spine are large enough to create pressures of several atmospheres. Most causes of excessive pressure in the skeletal system can be avoided by lifting properly and avoiding extreme physical activity. (See Forces and Torques in Muscles and Joints.)

There are many other interesting and medically significant pressures in the body. For example, pressure caused by various muscle actions drives food and waste through the digestive system. Stomach pressure behaves much like bladder pressure and is tied to the sensation of hunger. Pressure in the relaxed esophagus is normally negative because pressure in the chest cavity is normally negative. Positive pressure in the stomach may thus force acid into the esophagus, causing “heartburn.” Pressure in the middle ear can result in significant force on the eardrum if it differs greatly from atmospheric pressure, such as while scuba diving. The decrease in external pressure is also noticeable during plane flights (due to a decrease in the weight of air above

relative to that at the Earth's surface). The Eustachian tubes connect the middle ear to the throat and allow us to equalize pressure in the middle ear to avoid an imbalance of force on the eardrum.

Many pressures in the human body are associated with the flow of fluids. Fluid flow will be discussed in detail in the Fluid Dynamics and Its Biological and Medical Applications.

Section Summary

- Measuring blood pressure is among the most common of all medical examinations.
- The pressures in various parts of the body can be measured and often provide valuable medical indicators.
- The shape of the eye is maintained by fluid pressure, called intraocular pressure.
- When the circulation of fluid in the eye is blocked, it can lead to a buildup in pressure, a condition called glaucoma.
- Some of the other pressures in the body are spinal and skull pressures, bladder pressure, pressures in the skeletal system.

Problems & Exercises

Exercise:

Problem:

During forced exhalation, such as when blowing up a balloon, the diaphragm and chest muscles create a pressure of 60.0 mm Hg between the lungs and chest wall. What force in newtons does this pressure create on the 600 cm^2 surface area of the diaphragm?

Solution:

479 N

Exercise:

Problem:

You can chew through very tough objects with your incisors because they exert a large force on the small area of a pointed tooth. What pressure in pascals can you create by exerting a force of 500 N with your tooth on an area of 1.00 mm^2?

Exercise:

Problem:

One way to force air into an unconscious person's lungs is to squeeze on a balloon appropriately connected to the subject. What force must you exert on the balloon with your hands to create a gauge pressure of 4.00 cm water, assuming you squeeze on an effective area of 50.0 cm^2?

Solution:

1.96 N

Exercise:

Problem:

Heroes in movies hide beneath water and breathe through a hollow reed (villains never catch on to this trick). In practice, you cannot inhale in this manner if your lungs are more than 60.0 cm below the surface. What is the maximum negative gauge pressure you can create in your lungs on dry land, assuming you can achieve -3.00 cm water pressure with your lungs 60.0 cm below the surface?

Solution:

$-63.0\text{ cm H}_2\text{O}$

Exercise:

Problem:

Gauge pressure in the fluid surrounding an infant's brain may rise as high as 85.0 mm Hg (5 to 12 mm Hg is normal), creating an outward force large enough to make the skull grow abnormally large. (a) Calculate this outward force in newtons on each side of an infant's skull if the effective area of each side is 70.0 cm^2. (b) What is the net force acting on the skull?

Exercise:

Problem:

A full-term fetus typically has a mass of 3.50 kg. (a) What pressure does the weight of such a fetus create if it rests on the mother's bladder, supported on an area of 90.0 cm^2? (b) Convert this pressure to millimeters of mercury and determine if it alone is great enough to trigger the micturition reflex (it will add to any pressure already existing in the bladder).

Solution:

(a) $3.81 \times 10^3\text{ N/m}^2$

(b) 28.7 mm Hg, which is sufficient to trigger micturition reflex

Exercise:

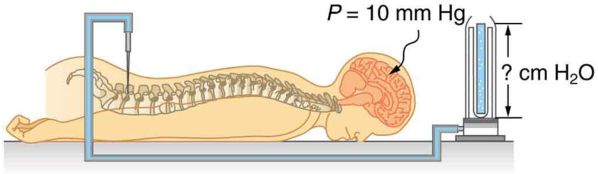
Problem:

If the pressure in the esophagus is -2.00 mm Hg while that in the stomach is $+20.0\text{ mm Hg}$, to what height could stomach fluid rise in the esophagus, assuming a density of 1.10 g/mL? (This movement will not occur if the muscle closing the lower end of the esophagus is working properly.)

Exercise:

Problem:

Pressure in the spinal fluid is measured as shown in [link]. If the pressure in the spinal fluid is 10.0 mm Hg: (a) What is the reading of the water manometer in cm water? (b) What is the reading if the person sits up, placing the top of the fluid 60 cm above the tap? The fluid density is 1.05 g/mL.



A water manometer used to measure pressure in the spinal fluid. The height of the fluid in the manometer is measured relative to the spinal column, and the manometer is open to the atmosphere. The measured pressure will be considerably greater if the person sits up.

Solution:

(a) 13.6 m water

(b) 76.5 cm water

Exercise:

Problem:

Calculate the maximum force in newtons exerted by the blood on an aneurysm, or ballooning, in a major artery, given the maximum blood pressure for this person is 150 mm Hg and the effective area of the aneurysm is $20.0\ \text{cm}^2$. Note that this force is great enough to cause further enlargement and subsequently greater force on the ever-thinner vessel wall.

Exercise:

Problem:

During heavy lifting, a disk between spinal vertebrae is subjected to a 5000-N compressional force. (a) What pressure is created, assuming that the disk has a uniform circular cross section 2.00 cm in radius? (b) What deformation is produced if the disk is 0.800 cm thick and has a Young's modulus of $1.5 \times 10^9\ \text{N/m}^2$?

Solution:

(a) $3.98 \times 10^6\ \text{Pa}$

(b) 2.1×10^{-3} cm

Exercise:

Problem:

When a person sits erect, increasing the vertical position of their brain by 36.0 cm, the heart must continue to pump blood to the brain at the same rate. (a) What is the gain in gravitational potential energy for 100 mL of blood raised 36.0 cm? (b) What is the drop in pressure, neglecting any losses due to friction? (c) Discuss how the gain in gravitational potential energy and the decrease in pressure are related.

Exercise:

Problem:

(a) How high will water rise in a glass capillary tube with a 0.500-mm radius? (b) How much gravitational potential energy does the water gain? (c) Discuss possible sources of this energy.

Solution:

(a) 2.97 cm

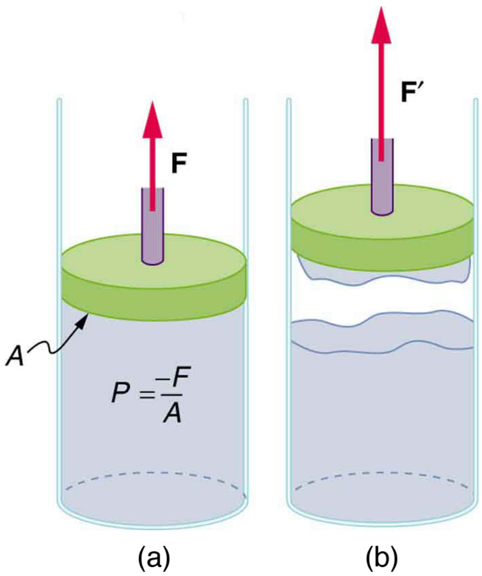
(b) 3.39×10^{-6} J

(c) Work is done by the surface tension force through an effective distance $h/2$ to raise the column of water.

Exercise:

Problem:

A negative pressure of 25.0 atm can sometimes be achieved with the device in [link] before the water separates. (a) To what height could such a negative gauge pressure raise water? (b) How much would a steel wire of the same diameter and length as this capillary stretch if suspended from above?



(a) When the piston is raised, it stretches the liquid slightly, putting it under tension and creating a negative absolute pressure $P = -F/A$ (b) The liquid eventually separates, giving an experimental limit to negative pressure in this liquid.

Exercise:

Problem:

Suppose you hit a steel nail with a 0.500-kg hammer, initially moving at $15.0\ \text{m/s}$ and brought to rest in 2.80 mm. (a) What average force is exerted on the nail? (b) How much is the nail compressed if it is 2.50 mm in diameter and 6.00-cm long? (c) What pressure is created on the 1.00-mm-diameter tip of the nail?

Solution:

(a) $2.01 \times 10^{4}\ \text{N}$

(b) $1.17 \times 10^{-3}\ \text{m}$

(c) $2.56 \times 10^{10}\ \text{N/m}^2$

Exercise:

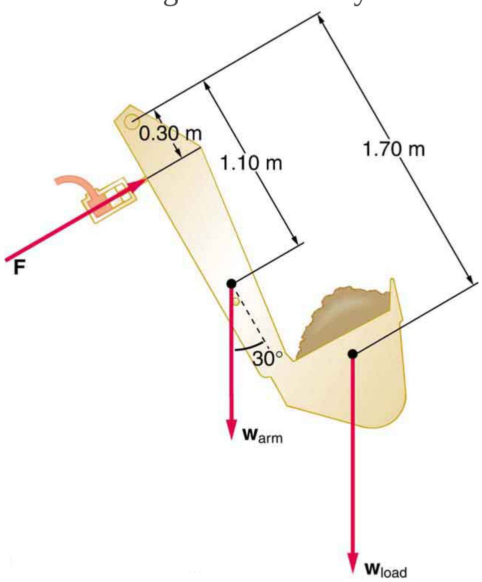
Problem:

Calculate the pressure due to the ocean at the bottom of the Marianas Trench near the Philippines, given its depth is 11.0 km and assuming the density of sea water is constant all the way down. (b) Calculate the percent decrease in volume of sea water due to such a pressure, assuming its bulk modulus is the same as water and is constant. (c) What would be the percent increase in its density? Is the assumption of constant density valid? Will the actual pressure be greater or smaller than that calculated under this assumption?

Exercise:

Problem:

The hydraulic system of a backhoe is used to lift a load as shown in [link]. (a) Calculate the force F the slave cylinder must exert to support the 400-kg load and the 150-kg brace and shovel. (b) What is the pressure in the hydraulic fluid if the slave cylinder is 2.50 cm in diameter? (c) What force would you have to exert on a lever with a mechanical advantage of 5.00 acting on a master cylinder 0.800 cm in diameter to create this pressure?



Hydraulic and mechanical lever systems are used in heavy machinery such as this back hoe.

Solution:

(a) $1.38 \times 10^4 \text{ N}$

(b) $2.81 \times 10^7 \text{ N/m}^2$

(c) 283 N

Exercise:

Problem:

Some miners wish to remove water from a mine shaft. A pipe is lowered to the water 90 m below, and a negative pressure is applied to raise the water. (a) Calculate the pressure needed to raise the water. (b) What is unreasonable about this pressure? (c) What is unreasonable about the premise?

Exercise:

Problem:

You are pumping up a bicycle tire with a hand pump, the piston of which has a 2.00-cm radius.

(a) What force in newtons must you exert to create a pressure of $6.90 \times 10^5 \text{ Pa}$ (b) What is unreasonable about this (a) result? (c) Which premises are unreasonable or inconsistent?

Solution:

(a) 867 N

(b) This is too much force to exert with a hand pump.

(c) The assumed radius of the pump is too large; it would be nearly two inches in diameter—too large for a pump or even a master cylinder. The pressure is reasonable for bicycle tires.

Exercise:

Problem:

Consider a group of people trying to stay afloat after their boat strikes a log in a lake. Construct a problem in which you calculate the number of people that can cling to the log and keep their heads out of the water. Among the variables to be considered are the size and density of the log, and what is needed to keep a person's head and arms above water without swimming or treading water.

Exercise:

Problem:

The alveoli in emphysema victims are damaged and effectively form larger sacs. Construct a problem in which you calculate the loss of pressure due to surface tension in the alveoli because of their larger average diameters. (Part of the lung's ability to expel air results from pressure created by surface tension in the alveoli.) Among the things to consider are the normal surface tension of the fluid lining the alveoli, the average alveolar radius in normal individuals and its average in emphysema sufferers.

Glossary

diastolic pressure
: minimum arterial blood pressure; indicator for the fluid balance

glaucoma
: condition caused by the buildup of fluid pressure in the eye

intraocular pressure
: fluid pressure in the eye

micturition reflex
: stimulates the feeling of needing to urinate, triggered by bladder pressure

systolic pressure
: maximum arterial blood pressure; indicator for the blood flow

Introduction to Work, Energy, and Energy Resources
class="introduction"

How many forms of energy can you identify in this photograph of a wind farm in Iowa? (credit: Jürgen from Sandesneben, Germany, Wikimedia Commons)

Energy plays an essential role both in everyday events and in scientific phenomena. You can no doubt name many forms of energy, from that provided by our foods, to the energy we use to run our cars, to the sunlight that warms us on the beach. You can also cite examples of what people call energy that may not be scientific, such as someone having an energetic personality. Not only does energy have many interesting forms, it is

involved in almost all phenomena, and is one of the most important concepts of physics. What makes it even more important is that the total amount of energy in the universe is constant. Energy can change forms, but it cannot appear from nothing or disappear without a trace. Energy is thus one of a handful of physical quantities that we say is *conserved*.

Conservation of energy (as physicists like to call the principle that energy can neither be created nor destroyed) is based on experiment. Even as scientists discovered new forms of energy, conservation of energy has always been found to apply. Perhaps the most dramatic example of this was supplied by Einstein when he suggested that mass is equivalent to energy (his famous equation $E = \mathrm{mc}^2$).

From a societal viewpoint, energy is one of the major building blocks of modern civilization. Energy resources are key limiting factors to economic growth. The world use of energy resources, especially oil, continues to grow, with ominous consequences economically, socially, politically, and environmentally. We will briefly examine the world's energy use patterns at the end of this chapter.

There is no simple, yet accurate, scientific definition for energy. Energy is characterized by its many forms and the fact that it is conserved. We can loosely define **energy** as the ability to do work, admitting that in some circumstances not all energy is available to do work. Because of the association of energy with work, we begin the chapter with a discussion of work. Work is intimately related to energy and how energy moves from one system to another or changes form.

Work: The Scientific Definition

- Explain how an object must be displaced for a force on it to do work.
- Explain how relative directions of force and displacement determine whether the work done is positive, negative, or zero.

What It Means to Do Work

The scientific definition of work differs in some ways from its everyday meaning. Certain things we think of as hard work, such as writing an exam or carrying a heavy load on level ground, are not work as defined by a scientist. The scientific definition of work reveals its relationship to energy—whenever work is done, energy is transferred.

For work, in the scientific sense, to be done, a force must be exerted and there must be displacement in the direction of the force.

Formally, the **work** done on a system by a constant force is defined to be *the product of the component of the force in the direction of motion times the distance through which the force acts*. For one-way motion in one dimension, this is expressed in equation form as
Equation:

$$W = \mid \mathbf{F} \mid (\cos\theta) \mid \mathbf{d} \mid,$$

where W is work, $\mathbf{d}$ is the displacement of the system, and θ is the angle between the force vector $\mathbf{F}$ and the displacement vector $\mathbf{d}$, as in [link]. We can also write this as
Equation:

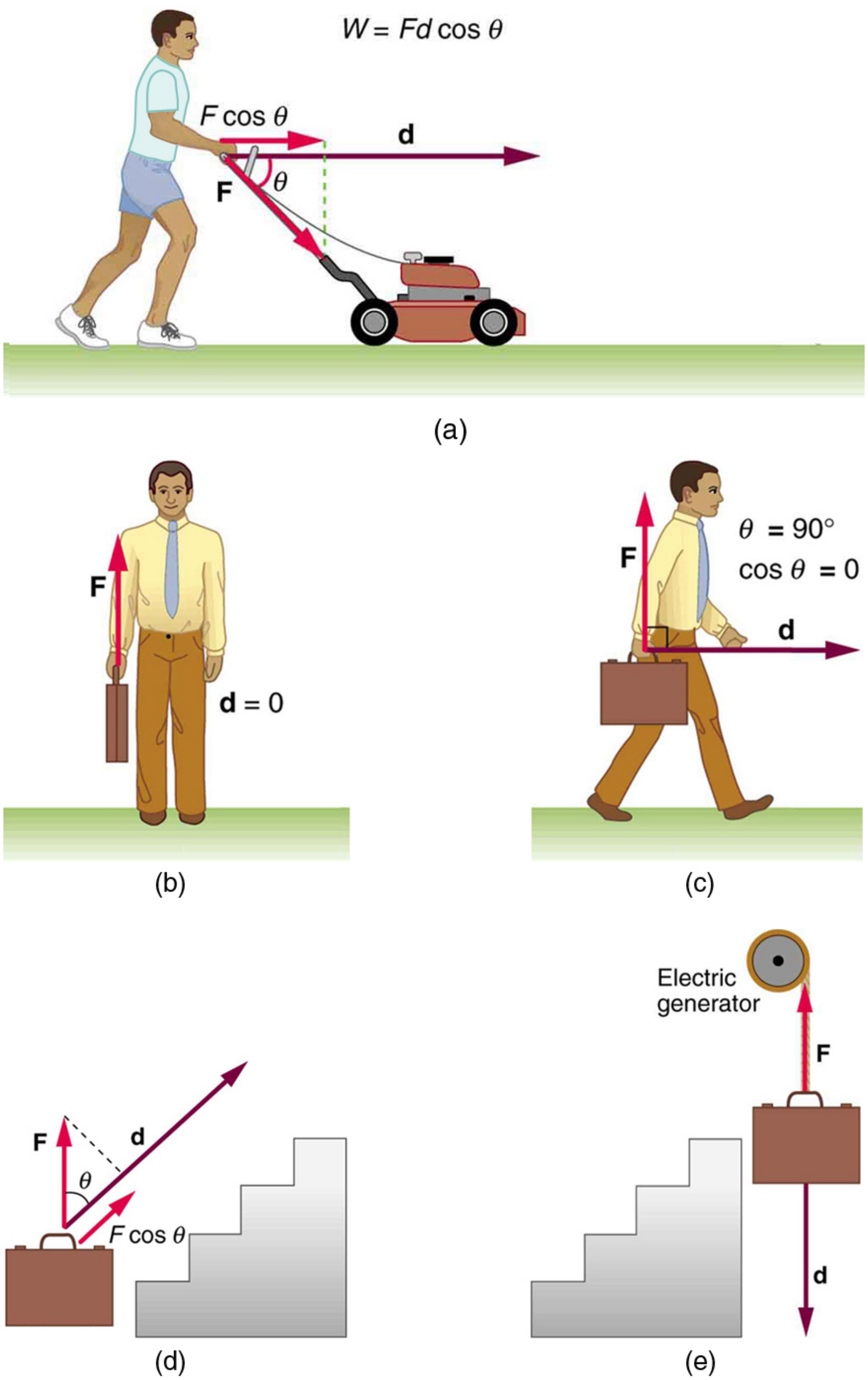
$$W = \mathrm{Fd}\ \cos\theta.$$

To find the work done on a system that undergoes motion that is not one-way or that is in two or three dimensions, we divide the motion into one-way one-dimensional segments and add up the work done over each segment.

Note:
What is Work?
The work done on a system by a constant force is *the product of the component of the force in the direction of motion times the distance through which the force acts*. For one-way motion in one dimension, this is expressed in equation form as
Equation:

$$W = \mathrm{Fd}\cos\theta,$$

where W is work, F is the magnitude of the force on the system, d is the magnitude of the displacement of the system, and θ is the angle between the force vector $\mathbf{F}$ and the displacement vector $\mathbf{d}$.



Examples of work. (a) The work done by the force $\mathbf{F}$ on this lawn mower is $\mathrm{Fd}\cos\theta$. Note that $F\cos\theta$ is the component of the force in the direction of motion. (b) A person holding a briefcase does no work on it, because there is no

displacement. No energy is transferred to or from the briefcase. (c) The person moving the briefcase horizontally at a constant speed does no work on it, and transfers no energy to it. (d) Work *is* done on the briefcase by carrying it up stairs at constant speed, because there is necessarily a component of force $\mathbf{F}$ in the direction of the motion. Energy is transferred to the briefcase and could in turn be used to do work. (e) When the briefcase is lowered, energy is transferred out of the briefcase and into an electric generator. Here the work done on the briefcase by the generator is negative, removing energy from the briefcase, because $\mathbf{F}$ and $\mathbf{d}$ are in opposite directions.

To examine what the definition of work means, let us consider the other situations shown in [link]. The person holding the briefcase in [link](b) does no work, for example. Here $d = 0$, so $W = 0$. Why is it you get tired just holding a load? The answer is that your muscles are doing work against one another, *but they are doing no work on the system of interest* (the “briefcase-Earth system”—see Gravitational Potential Energy for more details). There must be displacement for work to be done, and there must be a component of the force in the direction of the motion. For example, the person carrying the briefcase on level ground in [link](c) does no work on it, because the force is perpendicular to the motion. That is, $\cos 90^{\circ} = 0$, and so $W = 0$.

In contrast, when a force exerted on the system has a component in the direction of motion, such as in [link](d), work *is* done—energy is transferred to the briefcase. Finally, in [link](e), energy is transferred from the briefcase to a generator. There are two good ways to interpret this energy transfer. One interpretation is that the briefcase’s weight does work on the generator, giving it energy. The other interpretation is that the generator does negative work on the briefcase, thus removing energy from it. The drawing shows the latter, with the force from the generator upward

on the briefcase, and the displacement downward. This makes $\theta = 180^{\text{o}}$, and $\cos 180^{\text{o}} = -1$; therefore, W is negative.

Calculating Work

Work and energy have the same units. From the definition of work, we see that those units are force times distance. Thus, in SI units, work and energy are measured in **newton-meters**. A newton-meter is given the special name **joule** (J), and $1\text{ J} = 1\text{ N} \cdot \text{m} = 1\text{ kg} \cdot \text{m}^2/\text{s}^2$. One joule is not a large amount of energy; it would lift a small 100-gram apple a distance of about 1 meter.

Example:
Calculating the Work You Do to Push a Lawn Mower Across a Large Lawn

How much work is done on the lawn mower by the person in [link](a) if he exerts a constant force of 75.0 N at an angle 35^{o} below the horizontal and pushes the mower 25.0 m on level ground? Convert the amount of work from joules to kilocalories and compare it with this person's average daily intake of $10{,}000\text{ kJ}$ (about 2400 kcal) of food energy. One *calorie* (1 cal) of heat is the amount required to warm 1 g of water by 1^{o}C, and is equivalent to 4.184 J, while one *food calorie* (1 kcal) is equivalent to 4184 J.

Strategy

We can solve this problem by substituting the given values into the definition of work done on a system, stated in the equation $W = \text{Fd}\cos\theta$. The force, angle, and displacement are given, so that only the work W is unknown.

Solution

The equation for the work is

Equation:

$$W = \text{Fd}\cos\theta.$$

Substituting the known values gives

Equation:

$$\begin{aligned} W &= (75.0 \text{ N})(25.0 \text{ m}) \cos{(35.0^{\mathrm{o}})} \\ &= 1536 \text{ J} = 1.54 \times 10^{3} \text{ J}. \end{aligned}$$

Converting the work in joules to kilocalories yields $W = (1536 \text{ J})(1 \text{ kcal}/4184 \text{ J}) = 0.367 \text{ kcal}$. The ratio of the work done to the daily consumption is

Equation:

$$\frac{W}{2400 \text{ kcal}} = 1.53 \times 10^{-4}.$$

Discussion

This ratio is a tiny fraction of what the person consumes, but it is typical. Very little of the energy released in the consumption of food is used to do work. Even when we “work” all day long, less than 10% of our food energy intake is used to do work and more than 90% is converted to thermal energy or stored as chemical energy in fat.

Section Summary

- Work is the transfer of energy by a force acting on an object as it is displaced.
- The work W that a force $\mathbf{F}$ does on an object is the product of the magnitude F of the force, times the magnitude d of the displacement, times the cosine of the angle θ between them. In symbols,

  Equation:

  $$W = \text{Fd} \cos \theta.$$

- The SI unit for work and energy is the joule (J), where $1 \text{ J} = 1 \text{ N} \cdot \text{m} = 1 \text{ kg} \cdot \text{m}^2/\text{s}^2$.
- The work done by a force is zero if the displacement is either zero or perpendicular to the force.

- The work done is positive if the force and displacement have the same direction, and negative if they have opposite direction.

Conceptual Questions

Exercise:

Problem:

Give an example of something we think of as work in everyday circumstances that is not work in the scientific sense. Is energy transferred or changed in form in your example? If so, explain how this is accomplished without doing work.

Exercise:

Problem:

Give an example of a situation in which there is a force and a displacement, but the force does no work. Explain why it does no work.

Exercise:

Problem:

Describe a situation in which a force is exerted for a long time but does no work. Explain.

Problems & Exercises

Exercise:

Problem:

How much work does a supermarket checkout attendant do on a can of soup he pushes 0.600 m horizontally with a force of 5.00 N? Express your answer in joules and kilocalories.

Solution:

Equation:

$$3.00 \text{ J} = 7.17 \times 10^{-4} \text{ kcal}$$

Exercise:

Problem:

A 75.0-kg person climbs stairs, gaining 2.50 meters in height. Find the work done to accomplish this task.

Exercise:

Problem:

(a) Calculate the work done on a 1500-kg elevator car by its cable to lift it 40.0 m at constant speed, assuming friction averages 100 N. (b) What is the work done on the lift by the gravitational force in this process? (c) What is the total work done on the lift?

Solution:

(a) $5.92 \times 10^5 \text{ J}$

(b) $-5.88 \times 10^5 \text{ J}$

(c) The net force is zero.

Exercise:

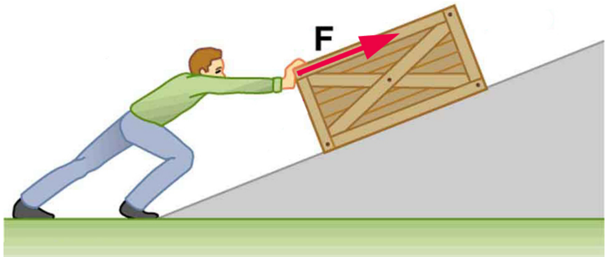
Problem:

Suppose a car travels 108 km at a speed of 30.0 m/s, and uses 2.0 gal of gasoline. Only 30% of the gasoline goes into useful work by the force that keeps the car moving at constant speed despite friction. (See [link] for the energy content of gasoline.) (a) What is the magnitude of the force exerted to keep the car moving at constant speed? (b) If the required force is directly proportional to speed, how many gallons will be used to drive 108 km at a speed of 28.0 m/s?

Exercise:

Problem:

Calculate the work done by an 85.0-kg man who pushes a crate 4.00 m up along a ramp that makes an angle of 20.0° with the horizontal. (See [link].) He exerts a force of 500 N on the crate parallel to the ramp and moves at a constant speed. Be certain to include the work he does on the crate *and* on his body to get up the ramp.



A man pushes a crate up a ramp.

Solution:

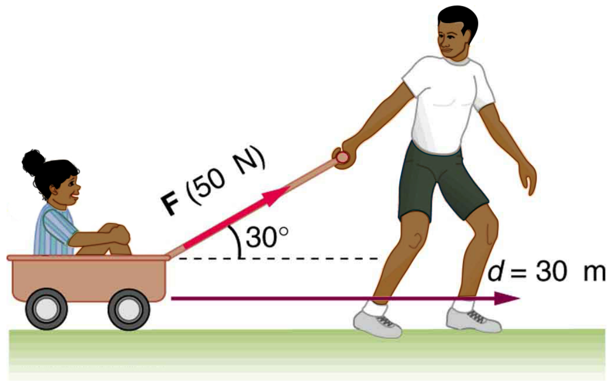
Equation:

$$3.14 \times 10^3 \text{ J}$$

Exercise:

Problem:

How much work is done by the boy pulling his sister 30.0 m in a wagon as shown in [link]? Assume no friction acts on the wagon.



The boy does work on the system of the wagon and the child when he pulls them as shown.

Exercise:

Problem:

A shopper pushes a grocery cart 20.0 m at constant speed on level ground, against a 35.0 N frictional force. He pushes in a direction 25.0° below the horizontal. (a) What is the work done on the cart by friction? (b) What is the work done on the cart by the gravitational force? (c) What is the work done on the cart by the shopper? (d) Find the force the shopper exerts, using energy considerations. (e) What is the total work done on the cart?

Solution:

(a) -700 J

(b) 0

(c) 700 J

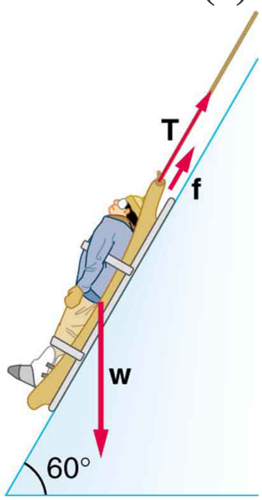
(d) 38.6 N

(e) 0

Exercise:

Problem:

Suppose the ski patrol lowers a rescue sled and victim, having a total mass of 90.0 kg, down a 60.0° slope at constant speed, as shown in [link]. The coefficient of friction between the sled and the snow is 0.100. (a) How much work is done by friction as the sled moves 30.0 m along the hill? (b) How much work is done by the rope on the sled in this distance? (c) What is the work done by the gravitational force on the sled? (d) What is the total work done?



A rescue sled and victim are lowered down a steep slope.

Glossary

energy

the ability to do work

work

the transfer of energy by a force that causes an object to be displaced; the product of the component of the force in the direction of the displacement and the magnitude of the displacement

joule

SI unit of work and energy, equal to one newton-meter

Kinetic Energy and the Work-Energy Theorem

- Explain work as a transfer of energy and net work as the work done by the net force.
- Explain and apply the work-energy theorem.

Work Transfers Energy

What happens to the work done on a system? Energy is transferred into the system, but in what form? Does it remain in the system or move on? The answers depend on the situation. For example, if the lawn mower in [link] (a) is pushed just hard enough to keep it going at a constant speed, then energy put into the mower by the person is removed continuously by friction, and eventually leaves the system in the form of heat transfer. In contrast, work done on the briefcase by the person carrying it up stairs in [link](d) is stored in the briefcase-Earth system and can be recovered at any time, as shown in [link](e). In fact, the building of the pyramids in ancient Egypt is an example of storing energy in a system by doing work on the system. Some of the energy imparted to the stone blocks in lifting them during construction of the pyramids remains in the stone-Earth system and has the potential to do work.

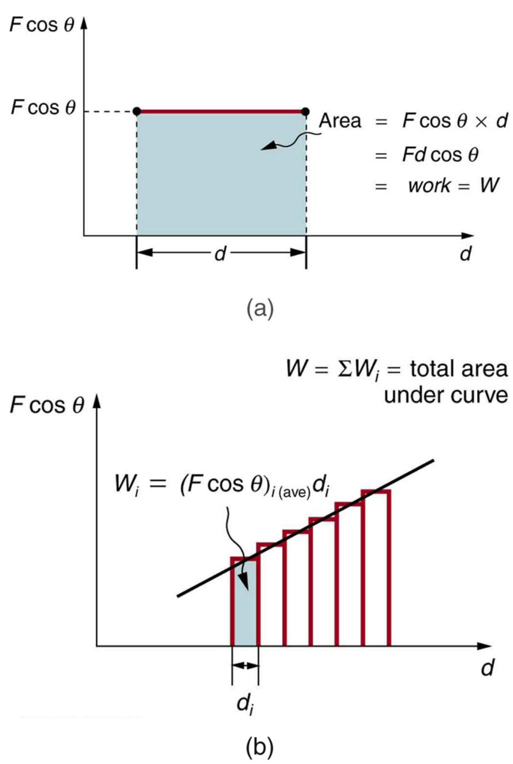
In this section we begin the study of various types of work and forms of energy. We will find that some types of work leave the energy of a system constant, for example, whereas others change the system in some way, such as making it move. We will also develop definitions of important forms of energy, such as the energy of motion.

Net Work and the Work-Energy Theorem

We know from the study of Newton’s laws in Dynamics: Force and Newton's Laws of Motion that net force causes acceleration. We will see in this section that work done by the net force gives a system energy of motion, and in the process we will also find an expression for the energy of motion.

Let us start by considering the total, or net, work done on a system. Net work is defined to be the sum of work done by all external forces—that is, **net work** is the work done by the net external force $\mathbf{F}_{\text{net}}$. In equation form, this is $W_{\text{net}} = F_{\text{net}} d \cos \theta$ where θ is the angle between the force vector and the displacement vector.

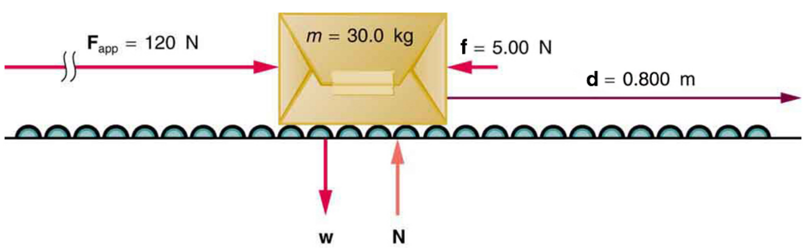
[link](a) shows a graph of force versus displacement for the component of the force in the direction of the displacement—that is, an $F \cos \theta$ vs. d graph. In this case, $F \cos \theta$ is constant. You can see that the area under the graph is $Fd \cos \theta$, or the work done. [link](b) shows a more general process where the force varies. The area under the curve is divided into strips, each having an average force $(F \cos \theta)_{i(\text{ave})}$. The work done is $(F \cos \theta)_{i(\text{ave})} d_i$ for each strip, and the total work done is the sum of the W_i. Thus the total work done is the total area under the curve, a useful property to which we shall refer later.



(a) A graph of $F \cos \theta$ vs. d, when $F \cos \theta$ is

constant. The area under the curve represents the work done by the force. (b) A graph of $F \cos \theta$ vs. d in which the force varies. The work done for each interval is the area of each strip; thus, the total area under the curve equals the total work done.

Net work will be simpler to examine if we consider a one-dimensional situation where a force is used to accelerate an object in a direction parallel to its initial velocity. Such a situation occurs for the package on the roller belt conveyor system shown in [link].



A package on a roller belt is pushed horizontally through a distance $\mathbf{d}$.

The force of gravity and the normal force acting on the package are perpendicular to the displacement and do no work. Moreover, they are also equal in magnitude and opposite in direction so they cancel in calculating the net force. The net force arises solely from the horizontal applied force $\mathbf{F}_{\text{app}}$ and the horizontal friction force $\mathbf{f}$. Thus, as expected, the net force is

parallel to the displacement, so that $\theta = 0^{\text{o}}$ and $\cos\theta = 1$, and the net work is given by

Equation:

$$W_{\text{net}} = F_{\text{net}}d.$$

The effect of the net force $\mathbf{F}_{\text{net}}$ is to accelerate the package from v_0 to v. The kinetic energy of the package increases, indicating that the net work done on the system is positive. (See [link].) By using Newton's second law, and doing some algebra, we can reach an interesting conclusion. Substituting $F_{\text{net}} = \text{ma}$ from Newton's second law gives

Equation:

$$W_{\text{net}} = \text{mad}.$$

To get a relationship between net work and the speed given to a system by the net force acting on it, we take $d = x - x_0$ and use the equation studied in Motion Equations for Constant Acceleration in One Dimension for the change in speed over a distance d if the acceleration has the constant value a; namely, $v^2 = {v_0}^2 + 2\text{ad}$ (note that a appears in the expression for the net work). Solving for acceleration gives $a = \frac{v^2 - {v_0}^2}{2d}$. When a is substituted into the preceding expression for W_{net}, we obtain

Equation:

$$W_{\text{net}} = m\left(\frac{v^2 - {v_0}^2}{2d}\right)d.$$

The d cancels, and we rearrange this to obtain

Equation:

$$W_{\text{ net}} = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2.$$

This expression is called the **work-energy theorem**, and it actually applies *in general* (even for forces that vary in direction and magnitude), although we have derived it for the special case of a constant force parallel to the displacement. The theorem implies that the net work on a system equals the change in the quantity $\frac{1}{2}mv^2$. This quantity is our first example of a form of energy.

Note:
The Work-Energy Theorem
The net work on a system equals the change in the quantity $\frac{1}{2}mv^2$.
Equation:

$$W_{\text{net}} = \frac{1}{2}mv^2 - \frac{1}{2}\text{mv}_0^2$$

The quantity $\frac{1}{2}mv^2$ in the work-energy theorem is defined to be the translational **kinetic energy** (KE) of a mass m moving at a speed v. (*Translational* kinetic energy is distinct from *rotational* kinetic energy, which is considered later.) In equation form, the translational kinetic energy,
Equation:

$$\text{KE} = \frac{1}{2}mv^2,$$

is the energy associated with translational motion. Kinetic energy is a form of energy associated with the motion of a particle, single body, or system of objects moving together.

We are aware that it takes energy to get an object, like a car or the package in [link], up to speed, but it may be a bit surprising that kinetic energy is proportional to speed squared. This proportionality means, for example, that a car traveling at 100 km/h has four times the kinetic energy it has at 50

km/h, helping to explain why high-speed collisions are so devastating. We will now consider a series of examples to illustrate various aspects of work and energy.

Example:
Calculating the Kinetic Energy of a Package

Suppose a 30.0-kg package on the roller belt conveyor system in [link] is moving at 0.500 m/s. What is its kinetic energy?

Strategy

Because the mass m and speed v are given, the kinetic energy can be calculated from its definition as given in the equation $\text{KE} = \frac{1}{2}mv^2$.

Solution

The kinetic energy is given by

Equation:

$$\text{KE} = \frac{1}{2}mv^2.$$

Entering known values gives

Equation:

$$\text{KE} = 0.5(30.0\text{ kg})(0.500\text{ m/s})^2,$$

which yields

Equation:

$$\text{KE} = 3.75\text{ kg}\cdot\text{m}^2/\text{s}^2 = 3.75\text{ J}.$$

Discussion

Note that the unit of kinetic energy is the joule, the same as the unit of work, as mentioned when work was first defined. It is also interesting that, although this is a fairly massive package, its kinetic energy is not large at this relatively low speed. This fact is consistent with the observation that people can move packages like this without exhausting themselves.

Example:

Determining the Work to Accelerate a Package

Suppose that you push on the 30.0-kg package in [link] with a constant force of 120 N through a distance of 0.800 m, and that the opposing friction force averages 5.00 N.

(a) Calculate the net work done on the package. (b) Solve the same problem as in part (a), this time by finding the work done by each force that contributes to the net force.

Strategy and Concept for (a)

This is a motion in one dimension problem, because the downward force (from the weight of the package) and the normal force have equal magnitude and opposite direction, so that they cancel in calculating the net force, while the applied force, friction, and the displacement are all horizontal. (See [link].) As expected, the net work is the net force times distance.

Solution for (a)

The net force is the push force minus friction, or $F_{\text{net}} = 120\text{ N} - 5.00\text{ N} = 115\text{ N}$. Thus the net work is

Equation:

$$\begin{aligned} W_{\text{net}} &= F_{\text{net}} d = (115\text{ N})(0.800\text{ m}) \\ &= 92.0\text{ N}\cdot\text{m} = 92.0\text{ J}. \end{aligned}$$

Discussion for (a)

This value is the net work done on the package. The person actually does more work than this, because friction opposes the motion. Friction does negative work and removes some of the energy the person expends and converts it to thermal energy. The net work equals the sum of the work done by each individual force.

Strategy and Concept for (b)

The forces acting on the package are gravity, the normal force, the force of friction, and the applied force. The normal force and force of gravity are each perpendicular to the displacement, and therefore do no work.

Solution for (b)

The applied force does work.

Equation:

$$\begin{aligned} W_{\text{app}} &= F_{\text{app}} d \cos(0^{\text{o}}) = F_{\text{app}} d \\ &= (120 \text{ N})(0.800 \text{ m}) \\ &= 96.0 \text{ J} \end{aligned}$$

The friction force and displacement are in opposite directions, so that $\theta = 180^{\text{o}}$, and the work done by friction is

Equation:

$$\begin{aligned} W_{\text{fr}} &= F_{\text{fr}} d \cos(180^{\text{o}}) = -F_{\text{fr}} d \\ &= -(5.00 \text{ N})(0.800 \text{ m}) \\ &= -4.00 \text{ J.} \end{aligned}$$

So the amounts of work done by gravity, by the normal force, by the applied force, and by friction are, respectively,

Equation:

$$\begin{aligned} W_{\text{gr}} &= 0, \\ W_{\text{N}} &= 0, \\ W_{\text{app}} &= 96.0 \text{ J}, \\ W_{\text{fr}} &= -4.00 \text{ J.} \end{aligned}$$

The total work done as the sum of the work done by each force is then seen to be

Equation:

$$W_{\text{total}} = W_{\text{gr}} + W_{\text{N}} + W_{\text{app}} + W_{\text{fr}} = 92.0 \text{ J.}$$

Discussion for (b)

The calculated total work W_{total} as the sum of the work by each force agrees, as expected, with the work W_{net} done by the net force. The work done by a collection of forces acting on an object can be calculated by either approach.

Example:

Determining Speed from Work and Energy

Find the speed of the package in [link] at the end of the push, using work and energy concepts.

Strategy

Here the work-energy theorem can be used, because we have just calculated the net work, W_{net}, and the initial kinetic energy, $\frac{1}{2}m{v_0}^2$. These calculations allow us to find the final kinetic energy, $\frac{1}{2}mv^2$, and thus the final speed v.

Solution

The work-energy theorem in equation form is

Equation:

$$W_{\text{net}} = \frac{1}{2}mv^2 - \frac{1}{2}m{v_0}^2.$$

Solving for $\frac{1}{2}mv^2$ gives

Equation:

$$\frac{1}{2}\text{mv}^2 = W_{\text{net}} + \frac{1}{2}m{v_0}^2.$$

Thus,

Equation:

$$\frac{1}{2}mv^2 = 92.0\text{ J} + 3.75\text{ J} = 95.75\text{ J}.$$

Solving for the final speed as requested and entering known values gives

Equation:

$$\begin{aligned} v &= \sqrt{\frac{2(95.75\text{ J})}{m}} = \sqrt{\frac{191.5\text{ kg}\cdot\text{m}^2/\text{s}^2}{30.0\text{ kg}}} \\ &= 2.53\text{ m/s}. \end{aligned}$$

Discussion

Using work and energy, we not only arrive at an answer, we see that the final kinetic energy is the sum of the initial kinetic energy and the net work

done on the package. This means that the work indeed adds to the energy of the package.

Example:
Work and Energy Can Reveal Distance, Too

How far does the package in [link] coast after the push, assuming friction remains constant? Use work and energy considerations.

Strategy

We know that once the person stops pushing, friction will bring the package to rest. In terms of energy, friction does negative work until it has removed all of the package's kinetic energy. The work done by friction is the force of friction times the distance traveled times the cosine of the angle between the friction force and displacement; hence, this gives us a way of finding the distance traveled after the person stops pushing.

Solution

The normal force and force of gravity cancel in calculating the net force. The horizontal friction force is then the net force, and it acts opposite to the displacement, so $\theta = 180º$. To reduce the kinetic energy of the package to zero, the work W_{fr} by friction must be minus the kinetic energy that the package started with plus what the package accumulated due to the pushing. Thus $W_{\text{fr}} = -95.75\text{ J}$. Furthermore, $W_{\text{fr}} = fd\prime \cos\theta = -fd\prime$, where $d\prime$ is the distance it takes to stop. Thus,

Equation:

$$d\prime = -\frac{W_{\text{fr}}}{f} = -\frac{-95.75\text{ J}}{5.00\text{ N}},$$

and so

Equation:

$$d\prime = 19.2\text{ m}.$$

Discussion

This is a reasonable distance for a package to coast on a relatively friction-free conveyor system. Note that the work done by friction is negative (the

force is in the opposite direction of motion), so it removes the kinetic energy.

Some of the examples in this section can be solved without considering energy, but at the expense of missing out on gaining insights about what work and energy are doing in this situation. On the whole, solutions involving energy are generally shorter and easier than those using kinematics and dynamics alone.

Section Summary

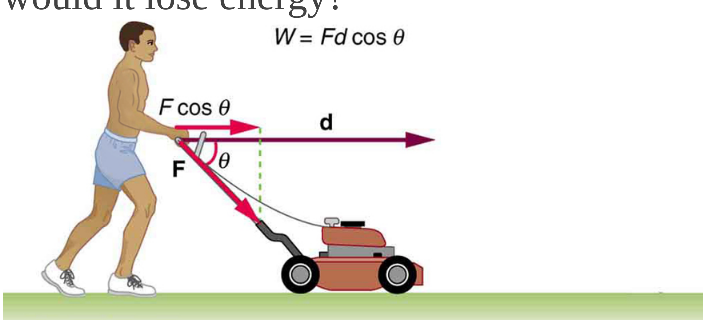
- The net work W_{net} is the work done by the net force acting on an object.
- Work done on an object transfers energy to the object.
- The translational kinetic energy of an object of mass m moving at speed v is $\text{KE} = \frac{1}{2}mv^2$.
- The work-energy theorem states that the net work W_{net} on a system changes its kinetic energy, $W_{\text{net}} = \frac{1}{2}mv^2 - \frac{1}{2}m{v_0}^2$.

Conceptual Questions

Exercise:

Problem:

The person in [link] does work on the lawn mower. Under what conditions would the mower gain energy? Under what conditions would it lose energy?



Exercise:

Problem:

Work done on a system puts energy into it. Work done by a system removes energy from it. Give an example for each statement.

Exercise:

Problem:

When solving for speed in [link], we kept only the positive root. Why?

Problems & Exercises

Exercise:

Problem:

Compare the kinetic energy of a 20,000-kg truck moving at 110 km/h with that of an 80.0-kg astronaut in orbit moving at 27,500 km/h.

Solution:

$1/250$

Exercise:

Problem:

(a) How fast must a 3000-kg elephant move to have the same kinetic energy as a 65.0-kg sprinter running at 10.0 m/s? (b) Discuss how the larger energies needed for the movement of larger animals would relate to metabolic rates.

Exercise:

Problem:

Confirm the value given for the kinetic energy of an aircraft carrier in [link]. You will need to look up the definition of a nautical mile (1 knot = 1 nautical mile/h).

Solution:

1.1×10^{10} J

Exercise:

Problem:

(a) Calculate the force needed to bring a 950-kg car to rest from a speed of 90.0 km/h in a distance of 120 m (a fairly typical distance for a non-panic stop). (b) Suppose instead the car hits a concrete abutment at full speed and is brought to a stop in 2.00 m. Calculate the force exerted on the car and compare it with the force found in part (a).

Exercise:

Problem:

A car's bumper is designed to withstand a 4.0-km/h (1.1-m/s) collision with an immovable object without damage to the body of the car. The bumper cushions the shock by absorbing the force over a distance. Calculate the magnitude of the average force on a bumper that collapses 0.200 m while bringing a 900-kg car to rest from an initial speed of 1.1 m/s.

Solution:

2.8×10^{3} N

Exercise:

Problem:

Boxing gloves are padded to lessen the force of a blow. (a) Calculate the force exerted by a boxing glove on an opponent's face, if the glove and face compress 7.50 cm during a blow in which the 7.00-kg arm and glove are brought to rest from an initial speed of 10.0 m/s. (b) Calculate the force exerted by an identical blow in the gory old days when no gloves were used and the knuckles and face would compress only 2.00 cm. (c) Discuss the magnitude of the force with glove on. Does it seem high enough to cause damage even though it is lower than the force with no glove?

Exercise:

Problem:

Using energy considerations, calculate the average force a 60.0-kg sprinter exerts backward on the track to accelerate from 2.00 to 8.00 m/s in a distance of 25.0 m, if he encounters a headwind that exerts an average force of 30.0 N against him.

Solution:

102 N

Glossary

net work
: work done by the net force, or vector sum of all the forces, acting on an object

work-energy theorem
: the result, based on Newton's laws, that the net work done on an object is equal to its change in kinetic energy

kinetic energy

the energy an object has by reason of its motion, equal to $\frac{1}{2}mv^2$ for the translational (i.e., non-rotational) motion of an object of mass m moving at speed v

Conservative Forces and Potential Energy

- Define conservative force, potential energy, and mechanical energy.
- Explain the potential energy of a spring in terms of its compression when Hooke's law applies.
- Use the work-energy theorem to show how having only conservative forces implies conservation of mechanical energy.

Potential Energy and Conservative Forces

Work is done by a force, and some forces, such as weight, have special characteristics. A **conservative force** is one, like the gravitational force, for which work done by or against it depends only on the starting and ending points of a motion and not on the path taken. We can define a **potential energy** (PE) for any conservative force, just as we did for the gravitational force. For example, when you wind up a toy, an egg timer, or an old-fashioned watch, you do work against its spring and store energy in it. (We treat these springs as ideal, in that we assume there is no friction and no production of thermal energy.) This stored energy is recoverable as work, and it is useful to think of it as potential energy contained in the spring. Indeed, the reason that the spring has this characteristic is that its force is *conservative*. That is, a conservative force results in stored or potential energy. Gravitational potential energy is one example, as is the energy stored in a spring. We will also see how conservative forces are related to the conservation of energy.

Note:
Potential Energy and Conservative Forces
Potential energy is the energy a system has due to position, shape, or configuration. It is stored energy that is completely recoverable.
A conservative force is one for which work done by or against it depends only on the starting and ending points of a motion and not on the path taken.
We can define a potential energy (PE) for any conservative force. The work done against a conservative force to reach a final configuration

depends on the configuration, not the path followed, and is the potential energy added.

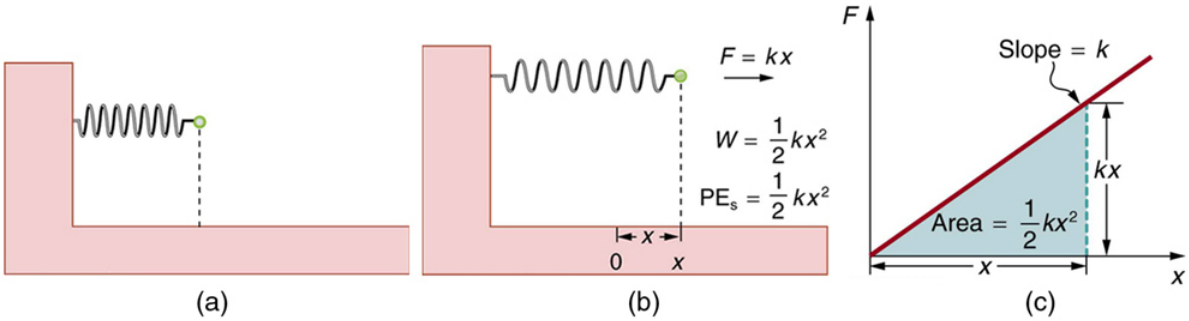
Potential Energy of a Spring

First, let us obtain an expression for the potential energy stored in a spring ($\mathrm{PE_s}$). We calculate the work done to stretch or compress a spring that obeys Hooke's law. (Hooke's law was examined in Elasticity: Stress and Strain, and states that the magnitude of force F on the spring and the resulting deformation ΔL are proportional, $F = k\Delta L$.) (See [link].) For our spring, we will replace ΔL (the amount of deformation produced by a force F) by the distance x that the spring is stretched or compressed along its length. So the force needed to stretch the spring has magnitude $\mathrm{F = kx}$, where k is the spring's force constant. The force increases linearly from 0 at the start to kx in the fully stretched position. The average force is $kx/2$. Thus the work done in stretching or compressing the spring is $W_\mathrm{s} = \mathrm{Fd} = \left(\frac{kx}{2}\right)x = \frac{1}{2}kx^2$. Alternatively, we noted in Kinetic Energy and the Work-Energy Theorem that the area under a graph of F vs. x is the work done by the force. In [link](c) we see that this area is also $\frac{1}{2}kx^2$. We therefore define the **potential energy of a spring**, $\mathrm{PE_s}$, to be

Equation:

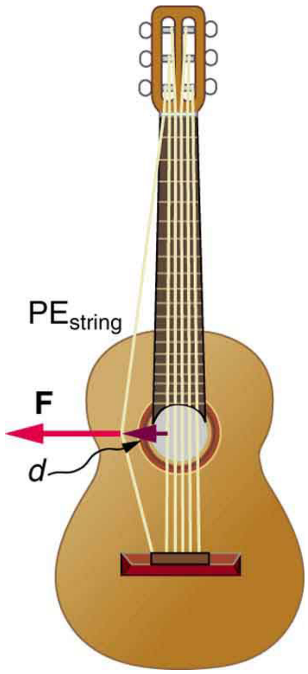
$$\mathrm{PE_s} = \frac{1}{2}\mathrm{kx}^2,$$

where k is the spring's force constant and x is the displacement from its undeformed position. The potential energy represents the work done *on* the spring and the energy stored in it as a result of stretching or compressing it a distance x. The potential energy of the spring $\mathrm{PE_s}$ does not depend on the path taken; it depends only on the stretch or squeeze x in the final configuration.



(a) An undeformed spring has no $\mathrm{PE_s}$ stored in it. (b) The force needed to stretch (or compress) the spring a distance x has a magnitude $F = kx$, and the work done to stretch (or compress) it is $\frac{1}{2}kx^2$. Because the force is conservative, this work is stored as potential energy $(\mathrm{PE_s})$ in the spring, and it can be fully recovered. (c) A graph of F vs. x has a slope of k, and the area under the graph is $\frac{1}{2}kx^2$. Thus the work done or potential energy stored is $\frac{1}{2}kx^2$.

The equation $\mathrm{PE_s} = \frac{1}{2}kx^2$ has general validity beyond the special case for which it was derived. Potential energy can be stored in any elastic medium by deforming it. Indeed, the general definition of **potential energy** is energy due to position, shape, or configuration. For shape or position deformations, stored energy is $\mathrm{PE_s} = \frac{1}{2}kx^2$, where k is the force constant of the particular system and x is its deformation. Another example is seen in [link] for a guitar string.



Work is done to deform the guitar string, giving it potential energy. When released, the potential energy is converted to kinetic energy and back to potential as the string oscillates back and forth. A very small fraction is dissipated as

sound
energy,
slowly
removing
energy from
the string.

Conservation of Mechanical Energy

Let us now consider what form the work-energy theorem takes when only conservative forces are involved. This will lead us to the conservation of energy principle. The work-energy theorem states that the net work done by all forces acting on a system equals its change in kinetic energy. In equation form, this is

Equation:

$$W_{\text{net}} = \frac{1}{2}mv^2 - \frac{1}{2}m{v_0}^2 = \Delta\text{KE}.$$

If only conservative forces act, then

Equation:

$$W_{\text{net}} = W_{\text{c}},$$

where W_{c} is the total work done by all conservative forces. Thus,

Equation:

$$W_{\text{c}} = \Delta\text{KE}.$$

Now, if the conservative force, such as the gravitational force or a spring force, does work, the system loses potential energy. That is, $W_{\text{c}} = -\Delta\text{PE}$. Therefore,

Equation:

$$-\Delta\text{PE} = \Delta\text{KE}$$

or
Equation:

$$\Delta\text{KE} + \Delta\text{PE} = 0.$$

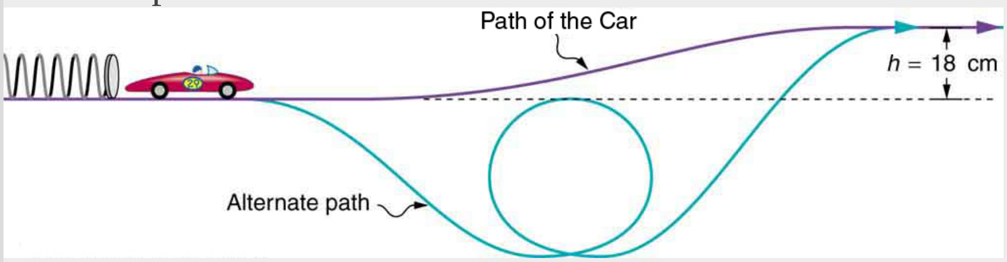
This equation means that the total kinetic and potential energy is constant for any process involving only conservative forces. That is,
Equation:

$$\left.\begin{array}{c} \text{KE} + \text{PE} = \text{constant} \\ \text{or} \\ \text{KE}_\text{i} + \text{PE}_\text{i} = \text{KE}_\text{f} + \text{PE}_\text{f} \end{array}\right\} \text{(conservative forces only)},$$

where i and f denote initial and final values. This equation is a form of the work-energy theorem for conservative forces; it is known as the **conservation of mechanical energy** principle. Remember that this applies to the extent that all the forces are conservative, so that friction is negligible. The total kinetic plus potential energy of a system is defined to be its **mechanical energy**, $(\text{KE} + \text{PE})$. In a system that experiences only conservative forces, there is a potential energy associated with each force, and the energy only changes form between KE and the various types of PE, with the total energy remaining constant.

Example:
Using Conservation of Mechanical Energy to Calculate the Speed of a Toy Car
A 0.100-kg toy car is propelled by a compressed spring, as shown in [link]. The car follows a track that rises 0.180 m above the starting point. The spring is compressed 4.00 cm and has a force constant of 250.0 N/m. Assuming work done by friction to be negligible, find (a) how fast the car

is going before it starts up the slope and (b) how fast it is going at the top of the slope.



A toy car is pushed by a compressed spring and coasts up a slope. Assuming negligible friction, the potential energy in the spring is first completely converted to kinetic energy, and then to a combination of kinetic and gravitational potential energy as the car rises. The details of the path are unimportant because all forces are conservative—the car would have the same final speed if it took the alternate path shown.

Strategy

The spring force and the gravitational force are conservative forces, so conservation of mechanical energy can be used. Thus,

Equation:

$$\mathrm{KE_i} + \mathrm{PE_i} = \mathrm{KE_f} + \mathrm{PE_f}$$

or

Equation:

$$\frac{1}{2}m{v_\mathrm{i}}^2 + mgh_\mathrm{i} + \frac{1}{2}k{x_\mathrm{i}}^2 = \frac{1}{2}m{v_\mathrm{f}}^2 + mgh_\mathrm{f} + \frac{1}{2}k{x_\mathrm{f}}^2,$$

where h is the height (vertical position) and x is the compression of the spring. This general statement looks complex but becomes much simpler when we start considering specific situations. First, we must identify the initial and final conditions in a problem; then, we enter them into the last equation to solve for an unknown.

Solution for (a)

This part of the problem is limited to conditions just before the car is released and just after it leaves the spring. Take the initial height to be zero, so that both h_{i} and h_{f} are zero. Furthermore, the initial speed v_{i} is zero and the final compression of the spring x_{f} is zero, and so several terms in the conservation of mechanical energy equation are zero and it simplifies to

Equation:

$$\frac{1}{2}k{x_{\mathrm{i}}}^2 = \frac{1}{2}m{v_{\mathrm{f}}}^2.$$

In other words, the initial potential energy in the spring is converted completely to kinetic energy in the absence of friction. Solving for the final speed and entering known values yields

Equation:

$$\begin{aligned} v_{\mathrm{f}} &= \sqrt{\frac{k}{m}}x_{\mathrm{i}} \\ &= \sqrt{\frac{250.0\ \mathrm{N/m}}{0.100\ \mathrm{kg}}}(0.0400\ \mathrm{m}) \\ &= 2.00\ \mathrm{m/s}. \end{aligned}$$

Solution for (b)

One method of finding the speed at the top of the slope is to consider conditions just before the car is released and just after it reaches the top of the slope, completely ignoring everything in between. Doing the same type of analysis to find which terms are zero, the conservation of mechanical energy becomes

Equation:

$$\frac{1}{2}\mathrm{k}{\mathrm{x}_{\mathrm{i}}}^2 = \frac{1}{2}\mathrm{m}{\mathrm{v}_{\mathrm{f}}}^2 + \mathrm{mgh}_{\mathrm{f}}.$$

This form of the equation means that the spring's initial potential energy is converted partly to gravitational potential energy and partly to kinetic energy. The final speed at the top of the slope will be less than at the bottom. Solving for v_{f} and substituting known values gives

Equation:

$$
\begin{aligned}
v_{\mathrm{f}} &= \sqrt{\frac{k{x_{\mathrm{i}}}^2}{m} - 2gh_{\mathrm{f}}} \\
&= \sqrt{\left(\frac{250.0\ \mathrm{N/m}}{0.100\ \mathrm{kg}}\right)(0.0400\ \mathrm{m})^2 - 2(9.80\ \mathrm{m/s^2})(0.180\ \mathrm{m})} \\
&= 0.687\ \mathrm{m/s}.
\end{aligned}
$$

Discussion

Another way to solve this problem is to realize that the car's kinetic energy before it goes up the slope is converted partly to potential energy—that is, to take the final conditions in part (a) to be the initial conditions in part (b).

Note that, for conservative forces, we do not directly calculate the work they do; rather, we consider their effects through their corresponding potential energies, just as we did in [link]. Note also that we do not consider details of the path taken—only the starting and ending points are important (as long as the path is not impossible). This assumption is usually a tremendous simplification, because the path may be complicated and forces may vary along the way.

Note:

PhET Explorations: Energy Skate Park

Learn about conservation of energy with a skater dude! Build tracks, ramps and jumps for the skater and view the kinetic energy, potential energy and friction as he moves. You can also take the skater to different planets or even space!

https://phet.colorado.edu/sims/html/energy-skate-park-basics/latest/energy-skate-park-basics_en.html

Section Summary

- A conservative force is one for which work depends only on the starting and ending points of a motion, not on the path taken.
- We can define potential energy (PE) for any conservative force, just as we defined $\mathrm{PE_g}$ for the gravitational force.
- The potential energy of a spring is $\mathrm{PE_s} = \frac{1}{2}kx^2$, where k is the spring's force constant and x is the displacement from its undeformed position.
- Mechanical energy is defined to be $\mathrm{KE} + \mathrm{PE}$ for a conservative force.
- When only conservative forces act on and within a system, the total mechanical energy is constant. In equation form,

Equation:

$$\mathrm{KE} + \mathrm{PE} = \text{constant}$$

or

$$\mathrm{KE_i} + \mathrm{PE_i} = \mathrm{KE_f} + \mathrm{PE_f}$$

where i and f denote initial and final values. This is known as the conservation of mechanical energy.

Conceptual Questions

Exercise:

Problem: What is a conservative force?

Exercise:

Problem:

The force exerted by a diving board is conservative, provided the internal friction is negligible. Assuming friction is negligible, describe changes in the potential energy of a diving board as a swimmer dives from it, starting just before the swimmer steps on the board until just after his feet leave it.

Exercise:

Problem:

Define mechanical energy. What is the relationship of mechanical energy to nonconservative forces? What happens to mechanical energy if only conservative forces act?

Exercise:

Problem:

What is the relationship of potential energy to conservative force?

Problems & Exercises

Exercise:

Problem:

A 5.00×10^5-kg subway train is brought to a stop from a speed of 0.500 m/s in 0.400 m by a large spring bumper at the end of its track. What is the force constant k of the spring?

Solution:
Equation:

$$7.81 \times 10^5 \text{ N/m}$$

Exercise:

Problem:

A pogo stick has a spring with a force constant of $2.50 \times 10^4 \text{ N/m}$, which can be compressed 12.0 cm. To what maximum height can a child jump on the stick using only the energy in the spring, if the child and stick have a total mass of 40.0 kg? Explicitly show how you follow the steps in the Problem-Solving Strategies for Energy.

Glossary

conservative force
: a force that does the same work for any given initial and final configuration, regardless of the path followed

potential energy
: energy due to position, shape, or configuration

potential energy of a spring
: the stored energy of a spring as a function of its displacement; when Hooke’s law applies, it is given by the expression $\frac{1}{2}kx^2$ where x is the distance the spring is compressed or extended and k is the spring constant

conservation of mechanical energy
: the rule that the sum of the kinetic energies and potential energies remains constant if only conservative forces act on and within a system

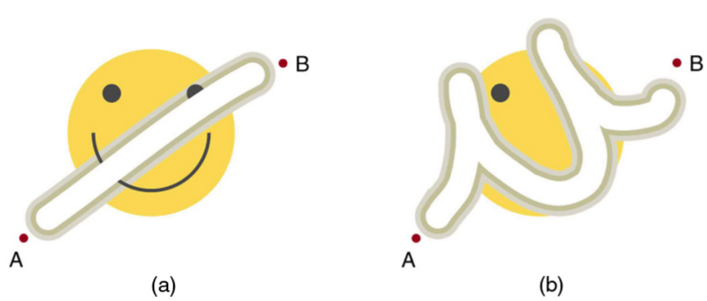
mechanical energy
: the sum of kinetic energy and potential energy

Nonconservative Forces

- Define nonconservative forces and explain how they affect mechanical energy.
- Show how the principle of conservation of energy can be applied by treating the conservative forces in terms of their potential energies and any nonconservative forces in terms of the work they do.

Nonconservative Forces and Friction

Forces are either conservative or nonconservative. Conservative forces were discussed in Conservative Forces and Potential Energy. A **nonconservative force** is one for which work depends on the path taken. Friction is a good example of a nonconservative force. As illustrated in [link], work done against friction depends on the length of the path between the starting and ending points. Because of this dependence on path, there is no potential energy associated with nonconservative forces. An important characteristic is that the work done by a nonconservative force *adds or removes mechanical energy from a system*. **Friction**, for example, creates **thermal energy** that dissipates, removing energy from the system. Furthermore, even if the thermal energy is retained or captured, it cannot be fully converted back to work, so it is lost or not recoverable in that sense as well.

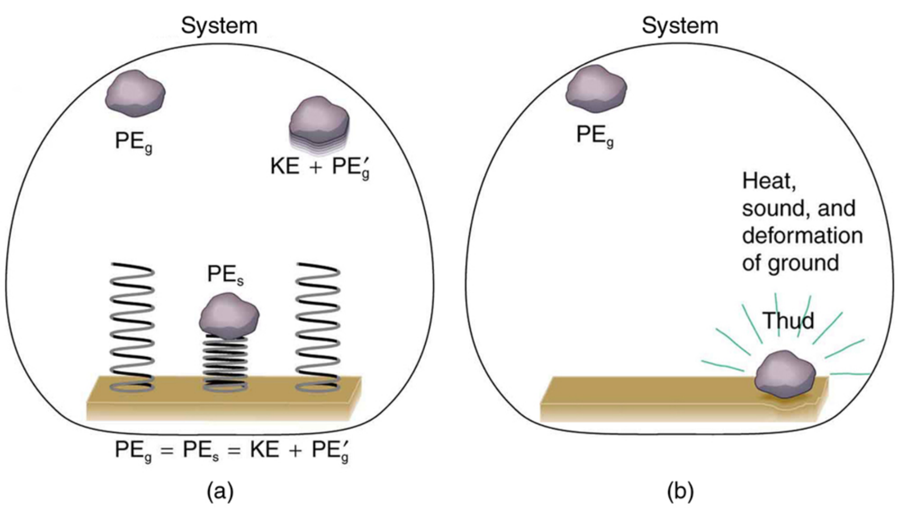


The amount of the happy face erased depends on the path taken by the eraser between points A and B, as does the work done against friction. Less work is done and less of the face

is erased for the path in (a) than for the path in (b). The force here is friction, and most of the work goes into thermal energy that subsequently leaves the system (the happy face plus the eraser). The energy expended cannot be fully recovered.

How Nonconservative Forces Affect Mechanical Energy

Mechanical energy *may* not be conserved when nonconservative forces act. For example, when a car is brought to a stop by friction on level ground, it loses kinetic energy, which is dissipated as thermal energy, reducing its mechanical energy. [link] compares the effects of conservative and nonconservative forces. We often choose to understand simpler systems such as that described in [link](a) first before studying more complicated systems as in [link](b).



Comparison of the effects of conservative and nonconservative forces on the mechanical energy of a system. (a) A system with only conservative

forces. When a rock is dropped onto a spring, its mechanical energy remains constant (neglecting air resistance) because the force in the spring is conservative. The spring can propel the rock back to its original height, where it once again has only potential energy due to gravity. (b) A system with nonconservative forces. When the same rock is dropped onto the ground, it is stopped by nonconservative forces that dissipate its mechanical energy as thermal energy, sound, and surface distortion. The rock has lost mechanical energy.

How the Work-Energy Theorem Applies

Now let us consider what form the work-energy theorem takes when both conservative and nonconservative forces act. We will see that the work done by nonconservative forces equals the change in the mechanical energy of a system. As noted in Kinetic Energy and the Work-Energy Theorem, the work-energy theorem states that the net work on a system equals the change in its kinetic energy, or $W_{\text{net}} = \Delta\text{KE}$. The net work is the sum of the work by nonconservative forces plus the work by conservative forces. That is,

Equation:

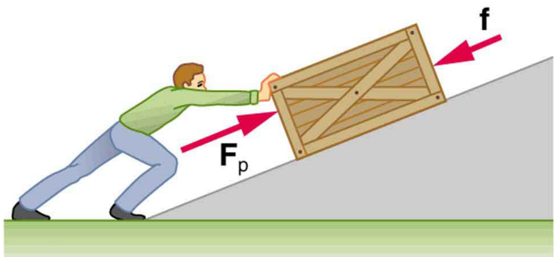
$$W_{\text{net}} = W_{\text{nc}} + W_{\text{c}},$$

so that

Equation:

$$W_{\text{nc}} + W_{\text{c}} = \Delta\text{KE},$$

where W_{nc} is the total work done by all nonconservative forces and W_{c} is the total work done by all conservative forces.



A person pushes a crate up a ramp, doing work on the crate. Friction and gravitational force (not shown) also do work on the crate; both forces oppose the person's push. As the crate is pushed up the ramp, it gains mechanical energy, implying that the work done by the person is greater than the work done by friction.

Consider [link], in which a person pushes a crate up a ramp and is opposed by friction. As in the previous section, we note that work done by a conservative force comes from a loss of gravitational potential energy, so that $W_{\text{c}} = -\Delta\text{PE}$. Substituting this equation into the previous one and solving for W_{nc} gives

Equation:

$$W_{\text{nc}} = \Delta\text{KE} + \Delta\text{PE}.$$

This equation means that the total mechanical energy $(\text{KE} + \text{PE})$ changes by exactly the amount of work done by nonconservative forces. In [link], this is the work done by the person minus the work done by friction. So even if energy is not conserved for the system of interest (such as the crate), we know that an equal amount of work was done to cause the change in total mechanical energy.

We rearrange $W_{\text{nc}} = \Delta\text{KE} + \Delta\text{PE}$ to obtain

Equation:

$$\text{KE}_{\text{i}} + \text{PE}_{\text{i}} + W_{\text{nc}} = \text{KE}_{\text{f}} + \text{PE}_{\text{f}}.$$

This means that the amount of work done by nonconservative forces adds to the mechanical energy of a system. If W_{nc} is positive, then mechanical energy is increased, such as when the person pushes the crate up the ramp in [link]. If W_{nc} is negative, then mechanical energy is decreased, such as when the rock hits the ground in [link](b). If W_{nc} is zero, then mechanical energy is conserved, and nonconservative forces are balanced. For example, when you push a lawn mower at constant speed on level ground, your work done is removed by the work of friction, and the mower has a constant energy.

Applying Energy Conservation with Nonconservative Forces

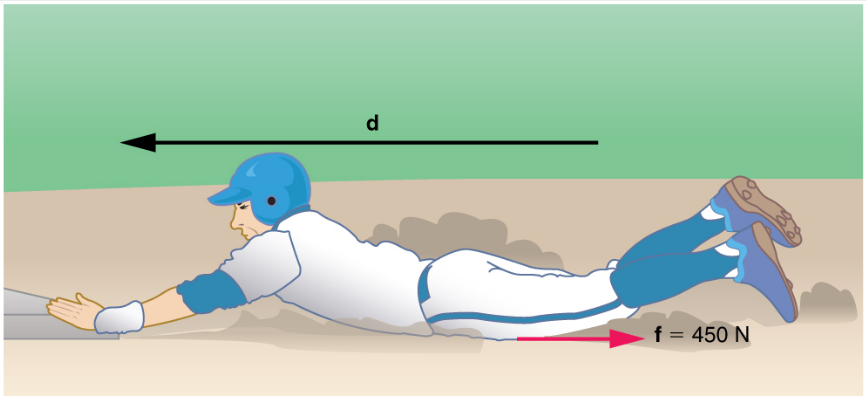
When no change in potential energy occurs, applying $\text{KE}_{\text{i}} + \text{PE}_{\text{i}} + W_{\text{nc}} = \text{KE}_{\text{f}} + \text{PE}_{\text{f}}$ amounts to applying the work-energy theorem by setting the change in kinetic energy to be equal to the net work done on the system, which in the most general case includes both conservative and nonconservative forces. But when seeking instead to find a change in total mechanical energy in situations that involve changes in both potential and kinetic energy, the previous equation $\text{KE}_{\text{i}} + \text{PE}_{\text{i}} + W_{\text{nc}} = \text{KE}_{\text{f}} + \text{PE}_{\text{f}}$ says that you can start by finding the change in mechanical energy that would have resulted from just the conservative forces, including the potential energy changes, and add to it the work done, with the proper sign, by any nonconservative forces involved.

Example:

Calculating Distance Traveled: How Far a Baseball Player Slides

Consider the situation shown in [link], where a baseball player slides to a stop on level ground. Using energy considerations, calculate the distance

the 65.0-kg baseball player slides, given that his initial speed is 6.00 m/s and the force of friction against him is a constant 450 N.



The baseball player slides to a stop in a distance *d*. In the process, friction removes the player's kinetic energy by doing an amount of work fd equal to the initial kinetic energy.

Strategy

Friction stops the player by converting his kinetic energy into other forms, including thermal energy. In terms of the work-energy theorem, the work done by friction, which is negative, is added to the initial kinetic energy to reduce it to zero. The work done by friction is negative, because **f** is in the opposite direction of the motion (that is, $\theta = 180^{\mathrm{o}}$, and so $\cos\theta = -1$). Thus $W_{\mathrm{nc}} = -\mathrm{fd}$. The equation simplifies to

Equation:

$$\frac{1}{2}m{v_{\mathrm{i}}}^2 - \mathrm{fd} = 0$$

or

Equation:

$$\mathrm{fd} = \frac{1}{2}m{v_{\mathrm{i}}}^2.$$

This equation can now be solved for the distance *d*.

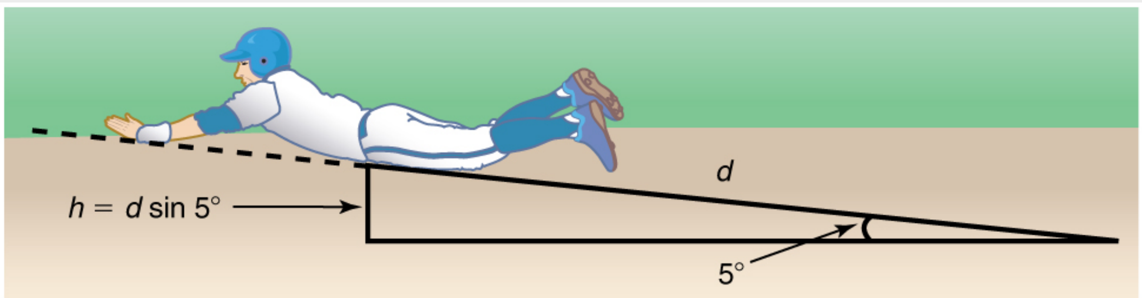
Solution

Solving the previous equation for d and substituting known values yields
Equation:

$$\begin{aligned} d &= \frac{m{v_\text{i}}^2}{2f} \\ &= \frac{(65.0\ \text{kg})(6.00\ \text{m/s})^2}{(2)(450\ \text{N})} \\ &= 2.60\ \text{m.} \end{aligned}$$

Discussion
The most important point of this example is that the amount of nonconservative work equals the change in mechanical energy. For example, you must work harder to stop a truck, with its large mechanical energy, than to stop a mosquito.

Example:
Calculating Distance Traveled: Sliding Up an Incline
Suppose that the player from [link] is running up a hill having a 5.00° incline upward with a surface similar to that in the baseball stadium. The player slides with the same initial speed, and the frictional force is still 450 N. Determine how far he slides.



The same baseball player slides to a stop on a 5.00° slope.

Strategy
In this case, the work done by the nonconservative friction force on the player reduces the mechanical energy he has from his kinetic energy at zero height, to the final mechanical energy he has by moving through

distance d to reach height h along the hill, with $h = d \sin 5.00^{\mathrm{o}}$. This is expressed by the equation

Equation:

$$\mathrm{KE_i} + \mathrm{PE_i} + W_{\mathrm{nc}} = \mathrm{KE_f} + \mathrm{PE_f}.$$

Solution

The work done by friction is again $W_{\mathrm{nc}} = -\mathrm{fd}$; initially the potential energy is $\mathrm{PE_i} = \mathrm{mg} \cdot 0 = 0$ and the kinetic energy is $\mathrm{KE_i} = \frac{1}{2}m{v_{\mathrm{i}}}^2$; the final energy contributions are $\mathrm{KE_f} = 0$ for the kinetic energy and $\mathrm{PE_f} = \mathrm{mgh} = \mathrm{mgd}\ \mathrm{sin}\ \theta$ for the potential energy.

Substituting these values gives

Equation:

$$\frac{1}{2}m{v_{\mathrm{i}}}^2 + 0 + \left(-fd\right) = 0 + mgd \sin \theta.$$

Solve this for d to obtain

Equation:

$$\begin{aligned} d &= \frac{\left(\frac{1}{2}\right)m{v_{\mathrm{i}}}^2}{f + mg \sin \theta} \\ &= \frac{(0.5)(65.0\ \mathrm{kg})(6.00\ \mathrm{m/s})^2}{450\ \mathrm{N} + (65.0\ \mathrm{kg})(9.80\ \mathrm{m/s^2})\ \sin\ (5.00^{\mathrm{o}})} \\ &= 2.31\ \mathrm{m.} \end{aligned}$$

Discussion

As might have been expected, the player slides a shorter distance by sliding uphill. Note that the problem could also have been solved in terms of the forces directly and the work energy theorem, instead of using the potential energy. This method would have required combining the normal force and force of gravity vectors, which no longer cancel each other because they point in different directions, and friction, to find the net force. You could then use the net force and the net work to find the distance d that reduces the kinetic energy to zero. By applying conservation of energy and using the potential energy instead, we need only consider the gravitational potential energy mgh, without combining and resolving force vectors. This simplifies the solution considerably.

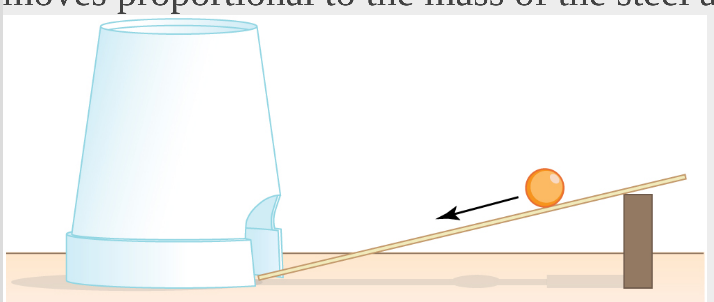
Note:

Making Connections: Take-Home Investigation—Determining Friction from the Stopping Distance

This experiment involves the conversion of gravitational potential energy into thermal energy. Use the ruler, book, and marble from Take-Home Investigation—Converting Potential to Kinetic Energy. In addition, you will need a foam cup with a small hole in the side, as shown in [link]. From the 10-cm position on the ruler, let the marble roll into the cup positioned at the bottom of the ruler. Measure the distance d the cup moves before stopping. What forces caused it to stop? What happened to the kinetic energy of the marble at the bottom of the ruler? Next, place the marble at the 20-cm and the 30-cm positions and again measure the distance the cup moves after the marble enters it. Plot the distance the cup moves versus the initial marble position on the ruler. Is this relationship linear?

With some simple assumptions, you can use these data to find the coefficient of kinetic friction μ_{k} of the cup on the table. The force of friction f on the cup is $\mu_{\text{k}}N$, where the normal force N is just the weight of the cup plus the marble. The normal force and force of gravity do no work because they are perpendicular to the displacement of the cup, which moves horizontally. The work done by friction is fd. You will need the mass of the marble as well to calculate its initial kinetic energy.

It is interesting to do the above experiment also with a steel marble (or ball bearing). Releasing it from the same positions on the ruler as you did with the glass marble, is the velocity of this steel marble the same as the velocity of the marble at the bottom of the ruler? Is the distance the cup moves proportional to the mass of the steel and glass marbles?



Rolling a marble down a ruler into a foam cup.

Note:
PhET Explorations: The Ramp
Explore forces, energy and work as you push household objects up and down a ramp. Lower and raise the ramp to see how the angle of inclination affects the parallel forces acting on the file cabinet. Graphs show forces, energy and work.

The Ramp

Section Summary

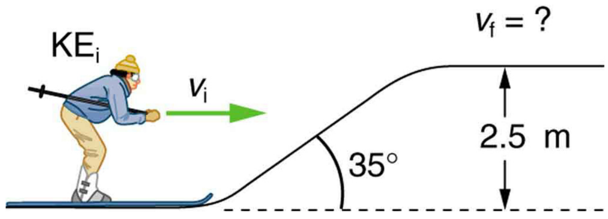
- A nonconservative force is one for which work depends on the path.
- Friction is an example of a nonconservative force that changes mechanical energy into thermal energy.
- Work W_{nc} done by a nonconservative force changes the mechanical energy of a system. In equation form, $W_{\text{nc}} = \Delta\text{KE} + \Delta\text{PE}$ or, equivalently, $\text{KE}_{\text{i}} + \text{PE}_{\text{i}} + W_{\text{nc}} = \text{KE}_{\text{f}} + \text{PE}_{\text{f}}$.
- When both conservative and nonconservative forces act, energy conservation can be applied and used to calculate motion in terms of the known potential energies of the conservative forces and the work done by nonconservative forces, instead of finding the net work from the net force, or having to directly apply Newton's laws.

Problems & Exercises

Exercise:

Problem:

A 60.0-kg skier with an initial speed of 12.0 m/s coasts up a 2.50-m-high rise as shown in [link]. Find her final speed at the top, given that the coefficient of friction between her skis and the snow is 0.0800. (Hint: Find the distance traveled up the incline assuming a straight-line path as shown in the figure.)



The skier's initial kinetic energy is partially used in coasting to the top of a rise.

Solution:

9.46 m/s

Exercise:

Problem:

(a) How high a hill can a car coast up (engine disengaged) if work done by friction is negligible and its initial speed is 110 km/h? (b) If, in actuality, a 750-kg car with an initial speed of 110 km/h is observed to coast up a hill to a height 22.0 m above its starting point, how much thermal energy was generated by friction? (c) What is the average force of friction if the hill has a slope 2.5° above the horizontal?

Glossary

nonconservative force

a force whose work depends on the path followed between the given initial and final configurations

friction

the force between surfaces that opposes one sliding on the other; friction changes mechanical energy into thermal energy

Conservation of Energy

- Explain the law of the conservation of energy.
- Describe some of the many forms of energy.
- Define efficiency of an energy conversion process as the fraction left as useful energy or work, rather than being transformed, for example, into thermal energy.

Law of Conservation of Energy

Energy, as we have noted, is conserved, making it one of the most important physical quantities in nature. The **law of conservation of energy** can be stated as follows:

Total energy is constant in any process. It may change in form or be transferred from one system to another, but the total remains the same.

We have explored some forms of energy and some ways it can be transferred from one system to another. This exploration led to the definition of two major types of energy—mechanical energy $(\mathrm{KE}+\mathrm{PE})$ and energy transferred via work done by nonconservative forces (W_{nc}). But energy takes *many* other forms, manifesting itself in *many* different ways, and we need to be able to deal with all of these before we can write an equation for the above general statement of the conservation of energy.

Other Forms of Energy than Mechanical Energy

At this point, we deal with all other forms of energy by lumping them into a single group called other energy (OE). Then we can state the conservation of energy in equation form as

Equation:

$$\mathrm{KE_i}+\mathrm{PE_i}+W_{\mathrm{nc}}+\mathrm{OE_i}=\mathrm{KE_f}+\mathrm{PE_f}+\mathrm{OE_f}.$$

All types of energy and work can be included in this very general statement of conservation of energy. Kinetic energy is KE, work done by a conservative force is represented by PE, work done by nonconservative forces is W_{nc}, and

all other energies are included as OE. This equation applies to all previous examples; in those situations OE was constant, and so it subtracted out and was not directly considered.

Note:
Making Connections: Usefulness of the Energy Conservation Principle
The fact that energy is conserved and has many forms makes it very important. You will find that energy is discussed in many contexts, because it is involved in all processes. It will also become apparent that many situations are best understood in terms of energy and that problems are often most easily conceptualized and solved by considering energy.

When does OE play a role? One example occurs when a person eats. Food is oxidized with the release of carbon dioxide, water, and energy. Some of this chemical energy is converted to kinetic energy when the person moves, to potential energy when the person changes altitude, and to thermal energy (another form of OE).

Some of the Many Forms of Energy

What are some other forms of energy? You can probably name a number of forms of energy not yet discussed. Many of these will be covered in later chapters, but let us detail a few here. **Electrical energy** is a common form that is converted to many other forms and does work in a wide range of practical situations. Fuels, such as gasoline and food, carry **chemical energy** that can be transferred to a system through oxidation. Chemical fuel can also produce electrical energy, such as in batteries. Batteries can in turn produce light, which is a very pure form of energy. Most energy sources on Earth are in fact stored energy from the energy we receive from the Sun. We sometimes refer to this as **radiant energy**, or electromagnetic radiation, which includes visible light, infrared, and ultraviolet radiation. **Nuclear energy** comes from processes that convert measurable amounts of mass into energy. Nuclear energy is transformed into the energy of sunlight, into electrical energy in power plants, and into the energy of the heat transfer and blast in weapons.

Atoms and molecules inside all objects are in random motion. This internal mechanical energy from the random motions is called **thermal energy**, because it is related to the temperature of the object. These and all other forms of energy can be converted into one another and can do work.

[link] gives the amount of energy stored, used, or released from various objects and in various phenomena. The range of energies and the variety of types and situations is impressive.

Note:
Problem-Solving Strategies for Energy
You will find the following problem-solving strategies useful whenever you deal with energy. The strategies help in organizing and reinforcing energy concepts. In fact, they are used in the examples presented in this chapter. The familiar general problem-solving strategies presented earlier—involving identifying physical principles, knowns, and unknowns, checking units, and so on—continue to be relevant here.

Step 1. Determine the system of interest and identify what information is given and what quantity is to be calculated. A sketch will help.

Step 2. Examine all the forces involved and determine whether you know or are given the potential energy from the work done by the forces. Then use step 3 or step 4.

Step 3. If you know the potential energies for the forces that enter into the problem, then forces are all conservative, and you can apply conservation of mechanical energy simply in terms of potential and kinetic energy. The equation expressing conservation of energy is

Equation:

$$\mathrm{KE_i} + \mathrm{PE_i} = \mathrm{KE_f} + \mathrm{PE_f}.$$

Step 4. If you know the potential energy for only some of the forces, possibly because some of them are nonconservative and do not have a potential energy, or if there are other energies that are not easily treated in terms of force and work, then the conservation of energy law in its most general form must be used.

Equation:

$$\mathrm{KE_i} + \mathrm{PE_i} + W_{\mathrm{nc}} + \mathrm{OE_i} = \mathrm{KE_f} + \mathrm{PE_f} + \mathrm{OE_f}.$$

In most problems, one or more of the terms is zero, simplifying its solution. Do not calculate W_{c}, the work done by conservative forces; it is already incorporated in the PE terms.
Step 5. You have already identified the types of work and energy involved (in step 2). Before solving for the unknown, *eliminate terms wherever possible* to simplify the algebra. For example, choose $h = 0$ at either the initial or final point, so that $\mathrm{PE_g}$ is zero there. Then solve for the unknown in the customary manner.
Step 6. *Check the answer to see if it is reasonable*. Once you have solved a problem, reexamine the forms of work and energy to see if you have set up the conservation of energy equation correctly. For example, work done against friction should be negative, potential energy at the bottom of a hill should be less than that at the top, and so on. Also check to see that the numerical value obtained is reasonable. For example, the final speed of a skateboarder who coasts down a 3-m-high ramp could reasonably be 20 km/h, but *not* 80 km/h.

Transformation of Energy

The transformation of energy from one form into others is happening all the time. The chemical energy in food is converted into thermal energy through metabolism; light energy is converted into chemical energy through photosynthesis. In a larger example, the chemical energy contained in coal is converted into thermal energy as it burns to turn water into steam in a boiler. This thermal energy in the steam in turn is converted to mechanical energy as it spins a turbine, which is connected to a generator to produce electrical energy. (In all of these examples, not all of the initial energy is converted into the forms mentioned. This important point is discussed later in this section.)

Another example of energy conversion occurs in a solar cell. Sunlight impinging on a solar cell (see [link]) produces electricity, which in turn can be used to run an electric motor. Energy is converted from the primary source of solar energy into electrical energy and then into mechanical energy.

Solar energy is converted into electrical energy by solar cells, which is used to run a motor in this solar-power aircraft. (credit: NASA)

Object/phenomenon	Energy in joules
Big Bang	10^{68}
Energy released in a supernova	10^{44}
Fusion of all the hydrogen in Earth’s oceans	10^{34}
Annual world energy use	4×10^{20}

Object/phenomenon	Energy in joules
Large fusion bomb (9 megaton)	3.8×10^{16}
1 kg hydrogen (fusion to helium)	6.4×10^{14}
1 kg uranium (nuclear fission)	8.0×10^{13}
Hiroshima-size fission bomb (10 kiloton)	4.2×10^{13}
90,000-ton aircraft carrier at 30 knots	1.1×10^{10}
1 barrel crude oil	5.9×10^{9}
1 ton TNT	4.2×10^{9}
1 gallon of gasoline	1.2×10^{8}
Daily home electricity use (developed countries)	7×10^{7}
Daily adult food intake (recommended)	1.2×10^{7}

Object/phenomenon	**Energy in joules**
1000-kg car at 90 km/h	3.1×10^5
1 g fat (9.3 kcal)	3.9×10^4
ATP hydrolysis reaction	3.2×10^4
1 g carbohydrate (4.1 kcal)	1.7×10^4
1 g protein (4.1 kcal)	1.7×10^4
Tennis ball at 100 km/h	22
Mosquito $\left(10^{-2}\text{ g at }0.5\text{ m/s}\right)$	1.3×10^{-6}
Single electron in a TV tube beam	4.0×10^{-15}
Energy to break one DNA strand	10^{-19}

Energy of Various Objects and Phenomena

Efficiency

Even though energy is conserved in an energy conversion process, the output of *useful energy* or work will be less than the energy input. The **efficiency** Eff of an energy conversion process is defined as

Equation:

$$\text{Efficiency(Eff)} = \frac{\text{useful energy or work output}}{\text{total energy input}} = \frac{W_{\text{out}}}{E_{\text{in}}}.$$

[link] lists some efficiencies of mechanical devices and human activities. In a coal-fired power plant, for example, about 40% of the chemical energy in the coal becomes useful electrical energy. The other 60% transforms into other (perhaps less useful) energy forms, such as thermal energy, which is then released to the environment through combustion gases and cooling towers.

Activity/device	**Efficiency (%)[footnote] Representative values**
Cycling and climbing	20
Swimming, surface	2
Swimming, submerged	4
Shoveling	3
Weightlifting	9
Steam engine	17
Gasoline engine	30

Activity/device	Efficiency (%)[footnote] Representative values
Diesel engine	35
Nuclear power plant	35
Coal power plant	42
Electric motor	98
Compact fluorescent light	20
Gas heater (residential)	90
Solar cell	10

Efficiency of the Human Body and Mechanical Devices

Note:
PhET Explorations: Masses and Springs
A realistic mass and spring laboratory. Hang masses from springs and adjust the spring stiffness and damping. You can even slow time. Transport the lab to different planets. A chart shows the kinetic, potential, and thermal energies for each spring.
https://phet.colorado.edu/sims/mass-spring-lab/mass-spring-lab_en.html

Section Summary

- The law of conservation of energy states that the total energy is constant in any process. Energy may change in form or be transferred from one system to another, but the total remains the same.
- When all forms of energy are considered, conservation of energy is written in equation form as

$\mathrm{KE_i} + \mathrm{PE_i} + W_{\mathrm{nc}} + \mathrm{OE_i} = \mathrm{KE_f} + \mathrm{PE_f} + \mathrm{OE_f}$, where OE is all **other forms of energy** besides mechanical energy.

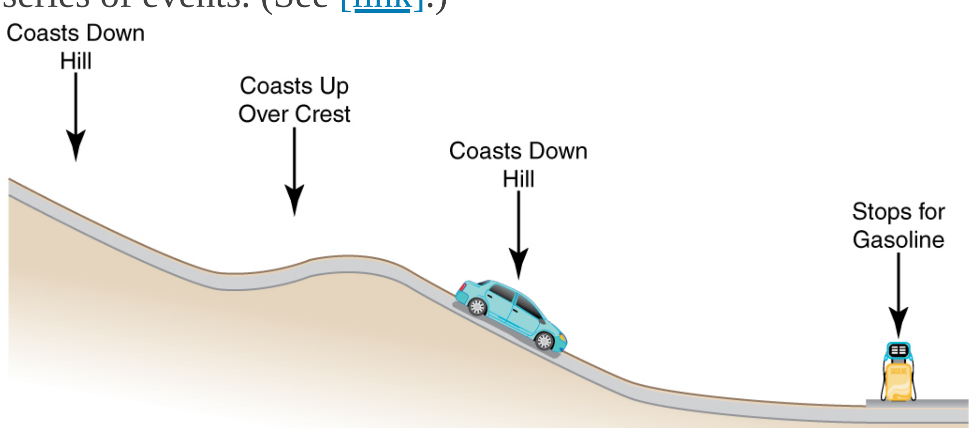
- Commonly encountered forms of energy include electric energy, chemical energy, radiant energy, nuclear energy, and thermal energy.
- Energy is often utilized to do work, but it is not possible to convert all the energy of a system to work.
- The efficiency Eff of a machine or human is defined to be $\mathrm{Eff} = \frac{W_{\mathrm{out}}}{E_{\mathrm{in}}}$, where W_{out} is useful work output and E_{in} is the energy consumed.

Conceptual Questions

Exercise:

Problem:

Consider the following scenario. A car for which friction is *not* negligible accelerates from rest down a hill, running out of gasoline after a short distance. The driver lets the car coast farther down the hill, then up and over a small crest. He then coasts down that hill into a gas station, where he brakes to a stop and fills the tank with gasoline. Identify the forms of energy the car has, and how they are changed and transferred in this series of events. (See [link].)



A car experiencing non-negligible friction coasts down a hill, over a small crest, then downhill again, and comes to a stop at a gas station.

Exercise:

Problem:

Describe the energy transfers and transformations for a javelin, starting from the point at which an athlete picks up the javelin and ending when the javelin is stuck into the ground after being thrown.

Exercise:

Problem:

Do devices with efficiencies of less than one violate the law of conservation of energy? Explain.

Exercise:

Problem:

List four different forms or types of energy. Give one example of a conversion from each of these forms to another form.

Exercise:

Problem: List the energy conversions that occur when riding a bicycle.

Problems & Exercises

Exercise:

Problem:

Using values from [link], how many DNA molecules could be broken by the energy carried by a single electron in the beam of an old-fashioned TV tube? (These electrons were not dangerous in themselves, but they did create dangerous x rays. Later model tube TVs had shielding that absorbed x rays before they escaped and exposed viewers.)

Solution:

4×10^4 molecules

Exercise:

Problem:

Using energy considerations and assuming negligible air resistance, show that a rock thrown from a bridge 20.0 m above water with an initial speed of 15.0 m/s strikes the water with a speed of 24.8 m/s independent of the direction thrown.

Solution:

Equating ΔPE_g and ΔKE, we obtain
$$v = \sqrt{2gh + {v_0}^2} = \sqrt{2(9.80\ \mathrm{m/s^2})(20.0\ \mathrm{m}) + (15.0\ \mathrm{m/s})^2} = 24.8\ \mathrm{m/s}$$

Exercise:

Problem:

If the energy in fusion bombs were used to supply the energy needs of the world, how many of the 9-megaton variety would be needed for a year's supply of energy (using data from [link])? This is not as far-fetched as it may sound—there are thousands of nuclear bombs, and their energy can be trapped in underground explosions and converted to electricity, as natural geothermal energy is.

Exercise:

Problem:

(a) Use of hydrogen fusion to supply energy is a dream that may be realized in the next century. Fusion would be a relatively clean and almost limitless supply of energy, as can be seen from [link]. To illustrate this, calculate how many years the present energy needs of the world could be supplied by one millionth of the oceans' hydrogen fusion energy. (b) How does this time compare with historically significant events, such as the duration of stable economic systems?

Solution:

(a) 25×10^6 years

(b) This is much, much longer than human time scales.

Glossary

law of conservation of energy
: the general law that total energy is constant in any process; energy may change in form or be transferred from one system to another, but the total remains the same

electrical energy
: the energy carried by a flow of charge

chemical energy
: the energy in a substance stored in the bonds between atoms and molecules that can be released in a chemical reaction

radiant energy
: the energy carried by electromagnetic waves

nuclear energy
: energy released by changes within atomic nuclei, such as the fusion of two light nuclei or the fission of a heavy nucleus

thermal energy
: the energy within an object due to the random motion of its atoms and molecules that accounts for the object's temperature

efficiency
: a measure of the effectiveness of the input of energy to do work; useful energy or work divided by the total input of energy

Power

- Calculate power by calculating changes in energy over time.
- Examine power consumption and calculations of the cost of energy consumed.

What is Power?

Power—the word conjures up many images: a professional football player muscling aside his opponent, a dragster roaring away from the starting line, a volcano blowing its lava into the atmosphere, or a rocket blasting off, as in [link].

This powerful rocket on the Space Shuttle *Endeavor* did work and consumed energy at a very high rate. (credit: NASA)

These images of power have in common the rapid performance of work, consistent with the scientific definition of **power** (P) as the rate at which work is done.

Note:
Power
Power is the rate at which work is done.
Equation:

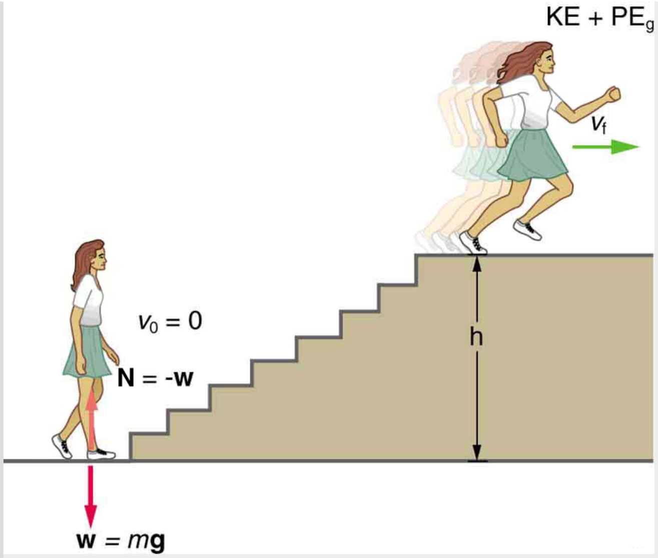
$$P = \frac{W}{t}$$

The SI unit for power is the **watt** (W), where 1 watt equals 1 joule/second $(1\ \mathrm{W} = 1\ \mathrm{J/s})$.

Because work is energy transfer, power is also the rate at which energy is expended. A 60-W light bulb, for example, expends 60 J of energy per second. Great power means a large amount of work or energy developed in a short time. For example, when a powerful car accelerates rapidly, it does a large amount of work and consumes a large amount of fuel in a short time.

Calculating Power from Energy

Example:
Calculating the Power to Climb Stairs
What is the power output for a 60.0-kg woman who runs up a 3.00 m high flight of stairs in 3.50 s, starting from rest but having a final speed of 2.00 m/s? (See [link].)



When this woman runs upstairs starting from rest, she converts the chemical energy originally from food into kinetic energy and gravitational potential energy. Her power output depends on how fast she does this.

Strategy and Concept

The work going into mechanical energy is $W = \mathrm{KE} + \mathrm{PE}$. At the bottom of the stairs, we take both KE and $\mathrm{PE_g}$ as initially zero; thus, $W = \mathrm{KE_f} + \mathrm{PE_g} = \frac{1}{2}m{v_\mathrm{f}}^2 + mgh$, where h is the vertical height of the stairs. Because all terms are given, we can calculate W and then divide it by time to get power.

Solution

Substituting the expression for W into the definition of power given in the previous equation, $P = W/t$ yields

Equation:

$$P = \frac{W}{t} = \frac{\frac{1}{2}m{v_\mathrm{f}}^2 + mgh}{t}.$$

Entering known values yields

Equation:

$$\begin{aligned} P &= \frac{0.5(60.0\text{ kg})(2.00\text{ m/s})^2 + (60.0\text{ kg})\left(9.80\text{ m/s}^2\right)(3.00\text{ m})}{3.50\text{ s}} \\ &= \frac{120\text{ J}+1764\text{ J}}{3.50\text{ s}} \\ &= 538\text{ W}. \end{aligned}$$

Discussion
The woman does 1764 J of work to move up the stairs compared with only 120 J to increase her kinetic energy; thus, most of her power output is required for climbing rather than accelerating.

It is impressive that this woman's useful power output is slightly less than 1 **horsepower** $(1\text{ hp} = 746\text{ W})$! People can generate more than a horsepower with their leg muscles for short periods of time by rapidly converting available blood sugar and oxygen into work output. (A horse can put out 1 hp for hours on end.) Once oxygen is depleted, power output decreases and the person begins to breathe rapidly to obtain oxygen to metabolize more food—this is known as the *aerobic* stage of exercise. If the woman climbed the stairs slowly, then her power output would be much less, although the amount of work done would be the same.

Note:
Making Connections: Take-Home Investigation—Measure Your Power Rating
Determine your own power rating by measuring the time it takes you to climb a flight of stairs. We will ignore the gain in kinetic energy, as the above example showed that it was a small portion of the energy gain. Don't expect that your output will be more than about 0.5 hp.

Examples of Power

Examples of power are limited only by the imagination, because there are as many types as there are forms of work and energy. (See [link] for some examples.) Sunlight reaching Earth's surface carries a maximum power of about 1.3 kilowatts per square meter $(\mathrm{kW/m^2})$. A tiny fraction of this is retained by Earth over the long term. Our consumption rate of fossil fuels is far greater than the rate at which they are stored, so it is inevitable that they will be depleted. Power implies that energy is transferred, perhaps changing form. It is never possible to change one form completely into another without losing some of it as thermal energy. For example, a 60-W incandescent bulb converts only 5 W of electrical power to light, with 55 W dissipating into thermal energy. Furthermore, the typical electric power plant converts only 35 to 40% of its fuel into electricity. The remainder becomes a huge amount of thermal energy that must be dispersed as heat transfer, as rapidly as it is created. A coal-fired power plant may produce 1000 megawatts; 1 megawatt (MW) is 10^6 W of electric power. But the power plant consumes chemical energy at a rate of about 2500 MW, creating heat transfer to the surroundings at a rate of 1500 MW. (See [link].)

Tremendous amounts of electric power are generated by coal-fired power plants such as this one in China, but an even larger amount of power goes into heat transfer to the surroundings.

The large cooling towers here are needed to transfer heat as rapidly as it is produced. The transfer of heat is not unique to coal plants but is an unavoidable consequence of generating electric power from any fuel—nuclear, coal, oil, natural gas, or the like. (credit: Kleinolive, Wikimedia Commons)

Object or Phenomenon	**Power in Watts**
Supernova (at peak)	5×10^{37}
Milky Way galaxy	10^{37}
Crab Nebula pulsar	10^{28}
The Sun	4×10^{26}

Object or Phenomenon	Power in Watts
Volcanic eruption (maximum)	4×10^{15}
Lightning bolt	2×10^{12}
Nuclear power plant (total electric and heat transfer)	3×10^{9}
Aircraft carrier (total useful and heat transfer)	10^{8}
Dragster (total useful and heat transfer)	2×10^{6}
Car (total useful and heat transfer)	8×10^{4}
Football player (total useful and heat transfer)	5×10^{3}
Clothes dryer	4×10^{3}
Person at rest (all heat transfer)	100

Object or Phenomenon	Power in Watts
Typical incandescent light bulb (total useful and heat transfer)	60
Heart, person at rest (total useful and heat transfer)	8
Electric clock	3
Pocket calculator	10^{-3}

Power Output or Consumption

Power and Energy Consumption

We usually have to pay for the energy we use. It is interesting and easy to estimate the cost of energy for an electrical appliance if its power consumption rate and time used are known. The higher the power consumption rate and the longer the appliance is used, the greater the cost of that appliance. The power consumption rate is $P = W/t = E/t$, where E is the energy supplied by the electricity company. So the energy consumed over a time t is

Equation:

$$E = \text{Pt}.$$

Electricity bills state the energy used in units of **kilowatt-hours** $(\text{kW} \cdot \text{h})$, which is the product of power in kilowatts and time in hours. This unit is convenient because electrical power consumption at the kilowatt level for hours at a time is typical.

Example:
Calculating Energy Costs
What is the cost of running a 0.200-kW computer 6.00 h per day for 30.0 d if the cost of electricity is \$0.120 per $\mathrm{kW \cdot h}$?

Strategy
Cost is based on energy consumed; thus, we must find E from $E = \mathrm{Pt}$ and then calculate the cost. Because electrical energy is expressed in $\mathrm{kW \cdot h}$, at the start of a problem such as this it is convenient to convert the units into kW and hours.

Solution
The energy consumed in $\mathrm{kW \cdot h}$ is

Equation:

$$\begin{aligned} E &= \mathrm{Pt} = (0.200\,\mathrm{kW})(6.00\,\mathrm{h/d})(30.0\,\mathrm{d}) \\ &= 36.0\,\mathrm{kW \cdot h}, \end{aligned}$$

and the cost is simply given by

Equation:

$$\text{cost} = (36.0\,\mathrm{kW \cdot h})(\$0.120\text{ per }\mathrm{kW \cdot h}) = \$4.32\text{ per month.}$$

Discussion
The cost of using the computer in this example is neither exorbitant nor negligible. It is clear that the cost is a combination of power and time. When both are high, such as for an air conditioner in the summer, the cost is high.

The motivation to save energy has become more compelling with its ever-increasing price. Armed with the knowledge that energy consumed is the product of power and time, you can estimate costs for yourself and make the necessary value judgments about where to save energy. Either power or time must be reduced. It is most cost-effective to limit the use of high-power devices that normally operate for long periods of time, such as water heaters and air conditioners. This would not include relatively high power devices like toasters, because they are on only a few minutes per day. It would also not include electric clocks, in spite of their 24-hour-per-day

usage, because they are very low power devices. It is sometimes possible to use devices that have greater efficiencies—that is, devices that consume less power to accomplish the same task. One example is the compact fluorescent light bulb, which produces over four times more light per watt of power consumed than its incandescent cousin.

Modern civilization depends on energy, but current levels of energy consumption and production are not sustainable. The likelihood of a link between global warming and fossil fuel use (with its concomitant production of carbon dioxide), has made reduction in energy use as well as a shift to non-fossil fuels of the utmost importance. Even though energy in an isolated system is a conserved quantity, the final result of most energy transformations is waste heat transfer to the environment, which is no longer useful for doing work. As we will discuss in more detail in Thermodynamics, the potential for energy to produce useful work has been "degraded" in the energy transformation.

Section Summary

- Power is the rate at which work is done, or in equation form, for the average power P for work W done over a time t, $P = W/t$.
- The SI unit for power is the watt (W), where $1\text{ W} = 1\text{ J/s}$.
- The power of many devices such as electric motors is also often expressed in horsepower (hp), where $1\text{ hp} = 746\text{ W}$.

Conceptual Questions

Exercise:

Problem:

Most electrical appliances are rated in watts. Does this rating depend on how long the appliance is on? (When off, it is a zero-watt device.) Explain in terms of the definition of power.

Exercise:

Problem:

Explain, in terms of the definition of power, why energy consumption is sometimes listed in kilowatt-hours rather than joules. What is the relationship between these two energy units?

Exercise:

Problem:

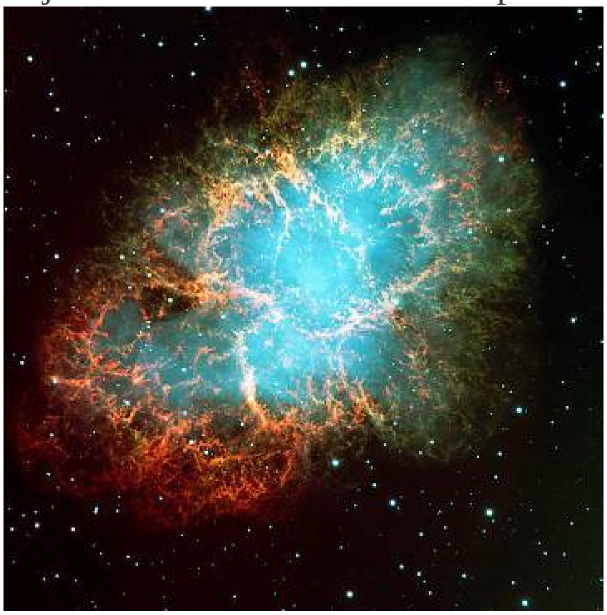
A spark of static electricity, such as that you might receive from a doorknob on a cold dry day, may carry a few hundred watts of power. Explain why you are not injured by such a spark.

Problems & Exercises

Exercise:

Problem:

The Crab Nebula (see [link]) pulsar is the remnant of a supernova that occurred in A.D. 1054. Using data from [link], calculate the approximate factor by which the power output of this astronomical object has declined since its explosion.



Crab Nebula (credit: ESO, via Wikimedia Commons)

Solution:
Equation:

$$2\times10^{-10}$$

Exercise:

Problem:

Suppose a star 1000 times brighter than our Sun (that is, emitting 1000 times the power) suddenly goes supernova. Using data from [link]: (a) By what factor does its power output increase? (b) How many times brighter than our entire Milky Way galaxy is the supernova? (c) Based on your answers, discuss whether it should be possible to observe supernovas in distant galaxies. Note that there are on the order of 10^{11} observable galaxies, the average brightness of which is somewhat less than our own galaxy.

Exercise:

Problem:

A person in good physical condition can put out 100 W of useful power for several hours at a stretch, perhaps by pedaling a mechanism that drives an electric generator. Neglecting any problems of generator efficiency and practical considerations such as resting time: (a) How many people would it take to run a 4.00-kW electric clothes dryer? (b) How many people would it take to replace a large electric power plant that generates 800 MW?

Solution:

(a) 40

(b) 8 million

Exercise:

Problem:

What is the cost of operating a 3.00-W electric clock for a year if the cost of electricity is $0.0900 per $\text{kW} \cdot \text{h}$?

Exercise:

Problem:

A large household air conditioner may consume 15.0 kW of power. What is the cost of operating this air conditioner 3.00 h per day for 30.0 d if the cost of electricity is $0.110 per $\text{kW} \cdot \text{h}$?

Solution:

$149

Exercise:

Problem:

(a) What is the average power consumption in watts of an appliance that uses $5.00\ \text{kW} \cdot \text{h}$ of energy per day? (b) How many joules of energy does this appliance consume in a year?

Exercise:

Problem:

(a) What is the average useful power output of a person who does 6.00×10^6 J of useful work in 8.00 h? (b) Working at this rate, how long will it take this person to lift 2000 kg of bricks 1.50 m to a platform? (Work done to lift his body can be omitted because it is not considered useful output here.)

Solution:

(a) 208 W

(b) 141 s

Exercise:

Problem:

A 500-kg dragster accelerates from rest to a final speed of 110 m/s in 400 m (about a quarter of a mile) and encounters an average frictional force of 1200 N. What is its average power output in watts and horsepower if this takes 7.30 s?

Exercise:

Problem:

(a) How long will it take an 850-kg car with a useful power output of 40.0 hp (1 hp = 746 W) to reach a speed of 15.0 m/s, neglecting friction? (b) How long will this acceleration take if the car also climbs a 3.00-m-high hill in the process?

Solution:

(a) 3.20 s

(b) 4.04 s

Exercise:

Problem:

(a) Find the useful power output of an elevator motor that lifts a 2500-kg load a height of 35.0 m in 12.0 s, if it also increases the speed from rest to 4.00 m/s. Note that the total mass of the counterbalanced system is 10,000 kg—so that only 2500 kg is raised in height, but the full 10,000 kg is accelerated. (b) What does it cost, if electricity is $0.0900 per $\text{kW} \cdot \text{h}$?

Exercise:

Problem:

(a) What is the available energy content, in joules, of a battery that operates a 2.00-W electric clock for 18 months? (b) How long can a battery that can supply 8.00×10^4 J run a pocket calculator that consumes energy at the rate of 1.00×10^{-3} W?

Solution:

(a) 9.46×10^7 J

(b) 2.54 y

Exercise:

Problem:

(a) How long would it take a 1.50×10^5-kg airplane with engines that produce 100 MW of power to reach a speed of 250 m/s and an altitude of 12.0 km if air resistance were negligible? (b) If it actually takes 900 s, what is the power? (c) Given this power, what is the average force of air resistance if the airplane takes 1200 s? (Hint: You must find the distance the plane travels in 1200 s assuming constant acceleration.)

Exercise:

Problem:

Calculate the power output needed for a 950-kg car to climb a 2.00^{o} slope at a constant 30.0 m/s while encountering wind resistance and friction totaling 600 N. Explicitly show how you follow the steps in the Problem-Solving Strategies for Energy.

Solution:

Identify knowns: $m = 950$ kg, slope angle $\theta = 2.00^{\text{o}}$, $v = 3.00$ m/s, $f = 600$ N

Identify unknowns: power P of the car, force F that car applies to road

Solve for unknown:

$$P = \frac{W}{t} = \frac{\text{Fd}}{t} = F\left(\frac{d}{t}\right) = \text{Fv},$$

where F is parallel to the incline and must oppose the resistive forces and the force of gravity:

$$F = f + w = 600\ \text{N} + \text{mg sin}\ \theta$$

Insert this into the expression for power and solve:

$$\begin{aligned} P &= (f + \text{mg sin}\ \theta)v \\ &= \left[600\ \text{N} + (950\ \text{kg})\left(9.80\ \text{m/s}^2\right)\text{sin}\ 2^{\text{o}}\right](30.0\ \text{m/s}) \\ &= 2.77 \times 10^4\ \text{W} \end{aligned}$$

About 28 kW (or about 37 hp) is reasonable for a car to climb a gentle incline.

Exercise:

Problem:

(a) Calculate the power per square meter reaching Earth's upper atmosphere from the Sun. (Take the power output of the Sun to be 4.00×10^{26} W.) (b) Part of this is absorbed and reflected by the atmosphere, so that a maximum of $1.30\ \text{kW/m}^2$ reaches Earth's surface. Calculate the area in km^2 of solar energy collectors needed to replace an electric power plant that generates 750 MW if the collectors convert an average of 2.00% of the maximum power into electricity. (This small conversion efficiency is due to the devices themselves, and the fact that the sun is directly overhead only briefly.) With the same assumptions, what area would be needed to meet the United States' energy needs $(1.05 \times 10^{20}\ \text{J})$? Australia's energy needs $(5.4 \times 10^{18}\ \text{J})$? China's energy needs $(6.3 \times 10^{19}\ \text{J})$? (These energy consumption values are from 2006.)

Glossary

power
: the rate at which work is done

watt
: (W) SI unit of power, with $1\ \mathrm{W} = 1\ \mathrm{J/s}$

horsepower
: an older non-SI unit of power, with $1\ \mathrm{hp} = 746\ \mathrm{W}$

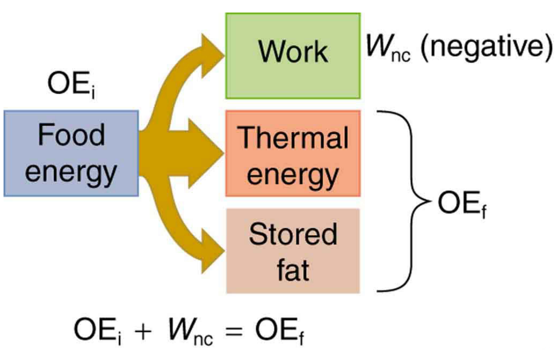
kilowatt-hour
: $(\mathrm{kW} \cdot \mathrm{h})$ unit used primarily for electrical energy provided by electric utility companies

Work, Energy, and Power in Humans

- Explain the human body’s consumption of energy when at rest vs. when engaged in activities that do useful work.
- Calculate the conversion of chemical energy in food into useful work.

Energy Conversion in Humans

Our own bodies, like all living organisms, are energy conversion machines. Conservation of energy implies that the chemical energy stored in food is converted into work, thermal energy, and/or stored as chemical energy in fatty tissue. (See [link].) The fraction going into each form depends both on how much we eat and on our level of physical activity. If we eat more than is needed to do work and stay warm, the remainder goes into body fat.



Energy consumed by humans is converted to work, thermal energy, and stored fat. By far the largest fraction goes to thermal energy, although the fraction varies depending on the type of physical activity.

Power Consumed at Rest

The *rate* at which the body uses food energy to sustain life and to do different activities is called the **metabolic rate**. The total energy conversion rate of a person *at rest* is called the **basal metabolic rate** (BMR) and is divided among various systems in the body, as shown in [link]. The largest fraction goes to the liver and spleen, with the brain coming next. Of course, during vigorous exercise, the energy consumption of the skeletal muscles and heart increase markedly. About 75% of the calories burned in a day go into these basic functions. The BMR is a function of age, gender, total body weight, and amount of muscle mass (which burns more calories than body fat). Athletes have a greater BMR due to this last factor.

Organ	**Power consumed at rest (W)**	**Oxygen consumption (mL/min)**	**Percent of BMR**
Liver & spleen	23	67	27
Brain	16	47	19
Skeletal muscle	15	45	18
Kidney	9	26	10
Heart	6	17	7
Other	16	48	19
Totals	85 W	250 mL/min	100%

Basal Metabolic Rates (BMR)

Energy consumption is directly proportional to oxygen consumption because the digestive process is basically one of oxidizing food. We can measure the energy people use during various activities by measuring their oxygen use. (See [link].) Approximately 20 kJ of energy are produced for each liter of oxygen consumed, independent of the type of food. [link] shows energy and oxygen consumption rates (power expended) for a variety of activities.

Power of Doing Useful Work

Work done by a person is sometimes called **useful work**, which is *work done on the outside world*, such as lifting weights. Useful work requires a force exerted through a distance on the outside world, and so it excludes internal work, such as that done by the heart when pumping blood. Useful work does include that done in climbing stairs or accelerating to a full run, because these are accomplished by exerting forces on the outside world. Forces exerted by the body are nonconservative, so that they can change the mechanical energy ($\mathrm{KE} + \mathrm{PE}$) of the system worked upon, and this is often the goal. A baseball player throwing a ball, for example, increases both the ball’s kinetic and potential energy.

If a person needs more energy than they consume, such as when doing vigorous work, the body must draw upon the chemical energy stored in fat. So exercise can be helpful in losing fat. However, the amount of exercise needed to produce a loss in fat, or to burn off extra calories consumed that day, can be large, as [link] illustrates.

Example:
Calculating Weight Loss from Exercising
If a person who normally requires an average of 12,000 kJ (3000 kcal) of food energy per day consumes 13,000 kJ per day, he will steadily gain weight. How much bicycling per day is required to work off this extra 1000 kJ?

Solution

[link] states that 400 W are used when cycling at a moderate speed. The time required to work off 1000 kJ at this rate is then

Equation:

$$\text{Time} = \frac{\text{energy}}{\left(\frac{\text{energy}}{\text{time}}\right)} = \frac{1000\ \text{kJ}}{400\ \text{W}} = 2500\ \text{s} = 42\ \text{min.}$$

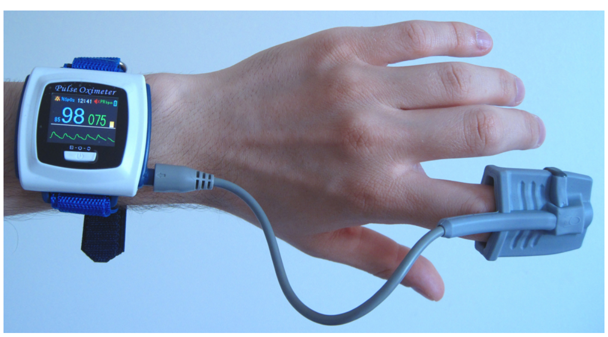
Discussion

If this person uses more energy than he or she consumes, the person's body will obtain the needed energy by metabolizing body fat. If the person uses 13,000 kJ but consumes only 12,000 kJ, then the amount of fat loss will be

Equation:

$$\text{Fat loss} = (1000\ \text{kJ})\left(\frac{1.0\ \text{g fat}}{39\ \text{kJ}}\right) = 26\ \text{g},$$

assuming the energy content of fat to be 39 kJ/g.



A pulse oxymeter is an apparatus that measures the amount of oxygen in blood. Oxymeters can be used to determine a person's metabolic rate, which is the rate at which food energy is converted to another form. Such

measurements can indicate the level of athletic conditioning as well as certain medical problems. (credit: UusiAjaja, Wikimedia Commons)

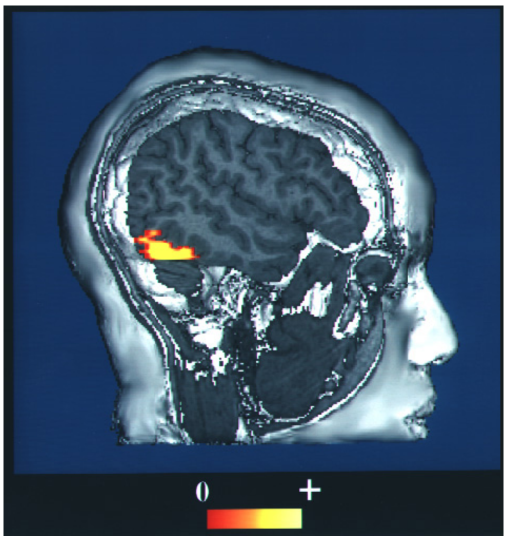
Activity	Energy consumption in watts	Oxygen consumption in liters O_2/min
Sleeping	83	0.24
Sitting at rest	120	0.34
Standing relaxed	125	0.36
Sitting in class	210	0.60
Walking (5 km/h)	280	0.80
Cycling (13–18 km/h)	400	1.14
Shivering	425	1.21
Playing tennis	440	1.26

Activity	Energy consumption in watts	Oxygen consumption in liters O_2/min
Swimming breaststroke	475	1.36
Ice skating (14.5 km/h)	545	1.56
Climbing stairs (116/min)	685	1.96
Cycling (21 km/h)	700	2.00
Running cross-country	740	2.12
Playing basketball	800	2.28
Cycling, professional racer	1855	5.30
Sprinting	2415	6.90

Energy and Oxygen Consumption Rates[footnote] (Power) for an average 76-kg male

All bodily functions, from thinking to lifting weights, require energy. (See [link].) The many small muscle actions accompanying all quiet activity, from sleeping to head scratching, ultimately become thermal energy, as do less visible muscle actions by the heart, lungs, and digestive tract. Shivering, in fact, is an involuntary response to low body temperature that pits muscles against one another to produce thermal energy in the body (and

do no work). The kidneys and liver consume a surprising amount of energy, but the biggest surprise of all is that a full 25% of all energy consumed by the body is used to maintain electrical potentials in all living cells. (Nerve cells use this electrical potential in nerve impulses.) This bioelectrical energy ultimately becomes mostly thermal energy, but some is utilized to power chemical processes such as in the kidneys and liver, and in fat production.



This fMRI scan shows an increased level of energy consumption in the vision center of the brain. Here, the patient was being asked to recognize faces. (credit: NIH via Wikimedia Commons)

Section Summary

- The human body converts energy stored in food into work, thermal energy, and/or chemical energy that is stored in fatty tissue.
- The *rate* at which the body uses food energy to sustain life and to do different activities is called the metabolic rate, and the corresponding rate when at rest is called the basal metabolic rate (BMR)
- The energy included in the basal metabolic rate is divided among various systems in the body, with the largest fraction going to the liver and spleen, and the brain coming next.
- About 75% of food calories are used to sustain basic body functions included in the basal metabolic rate.
- The energy consumption of people during various activities can be determined by measuring their oxygen use, because the digestive process is basically one of oxidizing food.

Conceptual Questions

Exercise:

Problem:

Explain why it is easier to climb a mountain on a zigzag path rather than one straight up the side. Is your increase in gravitational potential energy the same in both cases? Is your energy consumption the same in both?

Exercise:

Problem:

Do you do work on the outside world when you rub your hands together to warm them? What is the efficiency of this activity?

Exercise:

Problem:

Shivering is an involuntary response to lowered body temperature. What is the efficiency of the body when shivering, and is this a desirable value?

Exercise:

Problem:

Discuss the relative effectiveness of dieting and exercise in losing weight, noting that most athletic activities consume food energy at a rate of 400 to 500 W, while a single cup of yogurt can contain 1360 kJ (325 kcal). Specifically, is it likely that exercise alone will be sufficient to lose weight? You may wish to consider that regular exercise may increase the metabolic rate, whereas protracted dieting may reduce it.

Problems & Exercises

Exercise:

Problem:

(a) How long can you rapidly climb stairs (116/min) on the 93.0 kcal of energy in a 10.0-g pat of butter? (b) How many flights is this if each flight has 16 stairs?

Solution:

(a) 9.5 min

(b) 69 flights of stairs

Exercise:

Problem:

(a) What is the power output in watts and horsepower of a 70.0-kg sprinter who accelerates from rest to 10.0 m/s in 3.00 s? (b) Considering the amount of power generated, do you think a well-trained athlete could do this repetitively for long periods of time?

Exercise:

Problem:

Calculate the power output in watts and horsepower of a shot-putter who takes 1.20 s to accelerate the 7.27-kg shot from rest to 14.0 m/s, while raising it 0.800 m. (Do not include the power produced to accelerate his body.)

Shot putter at the Dornoch Highland Gathering in 2007. (credit: John Haslam, Flickr)

Solution:

641 W, 0.860 hp

Exercise:

Problem:

(a) What is the efficiency of an out-of-condition professor who does 2.10×10^5 J of useful work while metabolizing 500 kcal of food energy? (b) How many food calories would a well-conditioned athlete metabolize in doing the same work with an efficiency of 20%?

Exercise:

Problem:

Energy that is not utilized for work or heat transfer is converted to the chemical energy of body fat containing about 39 kJ/g. How many grams of fat will you gain if you eat 10,000 kJ (about 2500 kcal) one day and do nothing but sit relaxed for 16.0 h and sleep for the other 8.00 h? Use data from [link] for the energy consumption rates of these activities.

Solution:

31 g

Exercise:

Problem:

Using data from [link], calculate the daily energy needs of a person who sleeps for 7.00 h, walks for 2.00 h, attends classes for 4.00 h, cycles for 2.00 h, sits relaxed for 3.00 h, and studies for 6.00 h. (Studying consumes energy at the same rate as sitting in class.)

Exercise:

Problem:

What is the efficiency of a subject on a treadmill who puts out work at the rate of 100 W while consuming oxygen at the rate of 2.00 L/min? (Hint: See [link].)

Solution:

14.3%

Exercise:

Problem:

Shoveling snow can be extremely taxing because the arms have such a low efficiency in this activity. Suppose a person shoveling a footpath metabolizes food at the rate of 800 W. (a) What is her useful power output? (b) How long will it take her to lift 3000 kg of snow 1.20 m? (This could be the amount of heavy snow on 20 m of footpath.) (c) How much waste heat transfer in kilojoules will she generate in the process?

Exercise:

Problem:

Very large forces are produced in joints when a person jumps from some height to the ground. (a) Calculate the magnitude of the force produced if an 80.0-kg person jumps from a 0.600–m-high ledge and lands stiffly, compressing joint material 1.50 cm as a result. (Be certain to include the weight of the person.) (b) In practice the knees bend almost involuntarily to help extend the distance over which you stop. Calculate the magnitude of the force produced if the stopping distance is 0.300 m. (c) Compare both forces with the weight of the person.

Solution:

(a) 3.21×10^4 N

(b) 2.35×10^3 N

(c) Ratio of net force to weight of person is 41.0 in part (a); 3.00 in part (b)

Exercise:

Problem:

Jogging on hard surfaces with insufficiently padded shoes produces large forces in the feet and legs. (a) Calculate the magnitude of the force needed to stop the downward motion of a jogger’s leg, if his leg has a mass of 13.0 kg, a speed of 6.00 m/s, and stops in a distance of 1.50 cm. (Be certain to include the weight of the 75.0-kg jogger’s body.) (b) Compare this force with the weight of the jogger.

Exercise:

Problem:

(a) Calculate the energy in kJ used by a 55.0-kg woman who does 50 deep knee bends in which her center of mass is lowered and raised 0.400 m. (She does work in both directions.) You may assume her efficiency is 20%. (b) What is the average power consumption rate in watts if she does this in 3.00 min?

Solution:

(a) 108 kJ

(b) 599 W

Exercise:

Problem:

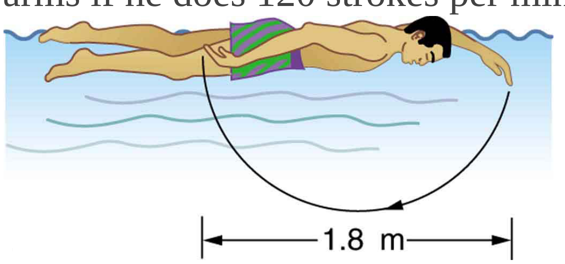
Kanellos Kanellopoulos flew 119 km from Crete to Santorini, Greece, on April 23, 1988, in the *Daedalus 88*, an aircraft powered by a bicycle-type drive mechanism (see [link]). His useful power output for the 234-min trip was about 350 W. Using the efficiency for cycling from [link], calculate the food energy in kilojoules he metabolized during the flight.

The Daedalus 88 in flight. (credit: NASA photo by Beasley)

Exercise:

Problem:

The swimmer shown in [link] exerts an average horizontal backward force of 80.0 N with his arm during each 1.80 m long stroke. (a) What is his work output in each stroke? (b) Calculate the power output of his arms if he does 120 strokes per minute.



Solution:

(a) 144 J

(b) 288 W

Exercise:

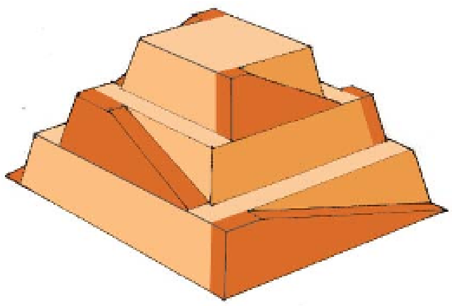
Problem:

Mountain climbers carry bottled oxygen when at very high altitudes. (a) Assuming that a mountain climber uses oxygen at twice the rate for climbing 116 stairs per minute (because of low air temperature and winds), calculate how many liters of oxygen a climber would need for 10.0 h of climbing. (These are liters at sea level.) Note that only 40% of the inhaled oxygen is utilized; the rest is exhaled. (b) How much useful work does the climber do if he and his equipment have a mass of 90.0 kg and he gains 1000 m of altitude? (c) What is his efficiency for the 10.0-h climb?

Exercise:

Problem:

The awe-inspiring Great Pyramid of Cheops was built more than 4500 years ago. Its square base, originally 230 m on a side, covered 13.1 acres, and it was 146 m high, with a mass of about 7×10^9 kg. (The pyramid's dimensions are slightly different today due to quarrying and some sagging.) Historians estimate that 20,000 workers spent 20 years to construct it, working 12-hour days, 330 days per year. (a) Calculate the gravitational potential energy stored in the pyramid, given its center of mass is at one-fourth its height. (b) Only a fraction of the workers lifted blocks; most were involved in support services such as building ramps (see [link]), bringing food and water, and hauling blocks to the site. Calculate the efficiency of the workers who did the lifting, assuming there were 1000 of them and they consumed food energy at the rate of 300 kcal/h. What does your answer imply about how much of their work went into block-lifting, versus how much work went into friction and lifting and lowering their own bodies? (c) Calculate the mass of food that had to be supplied each day, assuming that the average worker required 3600 kcal per day and that their diet was 5% protein, 60% carbohydrate, and 35% fat. (These proportions neglect the mass of bulk and nondigestible materials consumed.)



Ancient pyramids were probably constructed using ramps as simple machines. (credit: Franck Monnier, Wikimedia Commons)

Solution:

(a) 2.50×10^{12} J

(b) 2.52%

(c) 1.4×10^{4} kg (14 metric tons)

Exercise:

Problem:

(a) How long can you play tennis on the 800 kJ (about 200 kcal) of energy in a candy bar? (b) Does this seem like a long time? Discuss why exercise is necessary but may not be sufficient to cause a person to lose weight.

Glossary

metabolic rate
: the rate at which the body uses food energy to sustain life and to do different activities

basal metabolic rate
: the total energy conversion rate of a person at rest

useful work
: work done on an external system

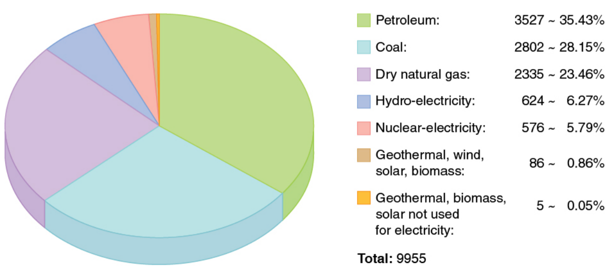
World Energy Use

- Describe the distinction between renewable and nonrenewable energy sources.
- Explain why the inevitable conversion of energy to less useful forms makes it necessary to conserve energy resources.

Energy is an important ingredient in all phases of society. We live in a very interdependent world, and access to adequate and reliable energy resources is crucial for economic growth and for maintaining the quality of our lives. But current levels of energy consumption and production are not sustainable. About 40% of the world's energy comes from oil, and much of that goes to transportation uses. Oil prices are dependent as much upon new (or foreseen) discoveries as they are upon political events and situations around the world. The U.S., with 4.5% of the world's population, consumes 24% of the world's oil production per year; 66% of that oil is imported!

Renewable and Nonrenewable Energy Sources

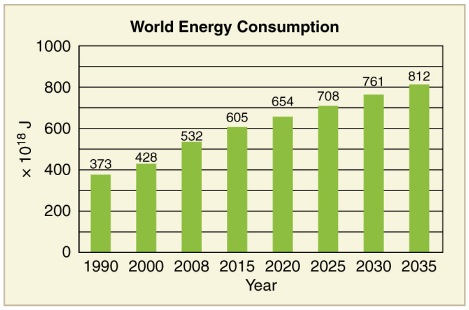
The principal energy resources used in the world are shown in [link]. The fuel mix has changed over the years but now is dominated by oil, although natural gas and solar contributions are increasing. **Renewable forms of energy** are those sources that cannot be used up, such as water, wind, solar, and biomass. About 85% of our energy comes from nonrenewable **fossil fuels**—oil, natural gas, coal. The likelihood of a link between global warming and fossil fuel use, with its production of carbon dioxide through combustion, has made, in the eyes of many scientists, a shift to non-fossil fuels of utmost importance—but it will not be easy.



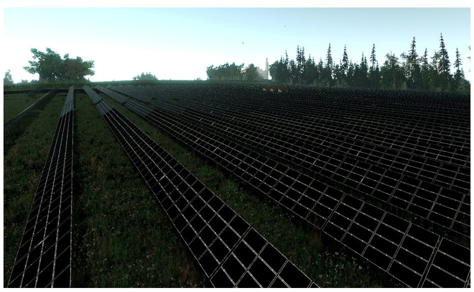
World energy consumption by source, in billions of kilowatt-hours: 2006. (credit: KVDP)

The World's Growing Energy Needs

World energy consumption continues to rise, especially in the developing countries. (See [link].) Global demand for energy has tripled in the past 50 years and might triple again in the next 30 years. While much of this growth will come from the rapidly booming economies of China and India, many of the developed countries, especially those in Europe, are hoping to meet their energy needs by expanding the use of renewable sources. Although presently only a small percentage, renewable energy is growing very fast, especially wind energy. For example, Germany plans to meet 20% of its electricity and 10% of its overall energy needs with renewable resources by the year 2020. (See [link].) Energy is a key constraint in the rapid economic growth of China and India. In 2003, China surpassed Japan as the world's second largest consumer of oil. However, over 1/3 of this is imported. Unlike most Western countries, coal dominates the commercial energy resources of China, accounting for 2/3 of its energy consumption. In 2009 China surpassed the United States as the largest generator of CO_2. In India, the main energy resources are biomass (wood and dung) and coal. Half of India's oil is imported. About 70% of India's electricity is generated by highly polluting coal. Yet there are sizeable strides being made in renewable energy. India has a rapidly growing wind energy base, and it has the largest solar cooking program in the world.



Past and projected world energy use (source: Based on data from U.S. Energy Information Administration, 2011)



Solar cell arrays at a power plant in Steindorf, Germany (credit: Michael Betke, Flickr)

[link] displays the 2006 commercial energy mix by country for some of the prime energy users in the world. While non-renewable sources dominate, some countries get a sizeable percentage of their electricity from renewable resources. For example, about 67% of New Zealand’s electricity demand is met by hydroelectric. Only 10% of the U.S. electricity is generated by renewable resources, primarily hydroelectric. It is difficult to determine total contributions of renewable energy in some countries with a large rural population, so these percentages in this table are left blank.

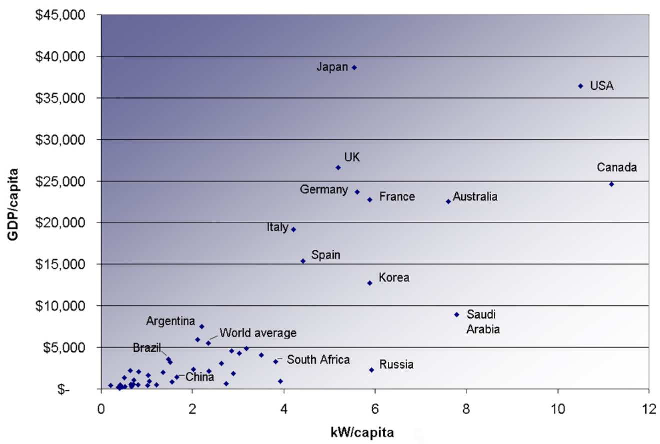
Country	Consumption, in EJ (10^{18} J)	Oil	Natural Gas	Coal	Nuclear	Hydro	Other Renewables
Australia	5.4	34%	17%	44%	0%	3%	1%

Country	Consumption, in EJ (10^{18} J)	Oil	Natural Gas	Coal	Nuclear	Hydro	Other Renewables
Brazil	9.6	48%	7%	5%	1%	35%	2%
China	63	22%	3%	69%	1%	6%	
Egypt	2.4	50%	41%	1%	0%	6%	
Germany	16	37%	24%	24%	11%	1%	3%
India	15	34%	7%	52%	1%	5%	
Indonesia	4.9	51%	26%	16%	0%	2%	3%
Japan	24	48%	14%	21%	12%	4%	1%
New Zealand	0.44	32%	26%	6%	0%	11%	19%
Russia	31	19%	53%	16%	5%	6%	
U.S.	105	40%	23%	22%	8%	3%	1%
World	**432**	**39%**	**23%**	**24%**	**6%**	**6%**	**2%**

Energy Consumption—Selected Countries (2006)

Energy and Economic Well-being

The last two columns in this table examine the energy and electricity use per capita. Economic well-being is dependent upon energy use, and in most countries higher standards of living, as measured by GDP (gross domestic product) per capita, are matched by higher levels of energy consumption per capita. This is borne out in [link]. Increased efficiency of energy use will change this dependency. A global problem is balancing energy resource development against the harmful effects upon the environment in its extraction and use.



Power consumption per capita versus GDP per capita for various countries. Note the increase in energy usage with increasing GDP. (2007, credit: Frank van Mierlo, Wikimedia Commons)

Conserving Energy

As we finish this chapter on energy and work, it is relevant to draw some distinctions between two sometimes misunderstood terms in the area of energy use. As has been mentioned elsewhere, the "law of the conservation of energy" is a very useful principle in analyzing physical processes. It is a statement that cannot be proven from basic principles, but is a very good bookkeeping device, and no exceptions have ever been found. It states that the total amount of energy in an isolated system will always remain constant. Related to this principle, but remarkably different from it, is the important philosophy of energy conservation. This concept has to do with seeking to decrease the amount of energy used by an individual or group through (1) reduced activities (e.g., turning down thermostats, driving fewer kilometers) and/or (2) increasing conversion efficiencies in the performance of a particular task—such as developing and using more efficient room heaters, cars that have greater miles-per-gallon ratings, energy-efficient compact fluorescent lights, etc.

Since energy in an isolated system is not destroyed or created or generated, one might wonder why we need to be concerned about our energy resources, since energy is a conserved quantity. The problem is that the final result of most energy transformations is waste heat transfer to the environment and conversion to energy forms no longer useful for doing work. To state it in another way, the potential for energy to produce useful work has been "degraded" in the energy transformation. (This will be discussed in more detail in Thermodynamics.)

Section Summary

- The relative use of different fuels to provide energy has changed over the years, but fuel use is currently dominated by oil, although natural gas and solar contributions are increasing.
- Although non-renewable sources dominate, some countries meet a sizeable percentage of their electricity needs from renewable resources.
- The United States obtains only about 10% of its energy from renewable sources, mostly hydroelectric power.
- Economic well-being is dependent upon energy use, and in most countries higher standards of living, as measured by GDP (Gross Domestic Product) per capita, are matched by higher levels of energy consumption per capita.
- Even though, in accordance with the law of conservation of energy, energy can never be created or destroyed, energy that can be used to do work is always partly converted to less useful forms, such as waste heat to the environment, in all of our uses of energy for practical purposes.

Conceptual Questions

Exercise:

Problem:

What is the difference between energy conservation and the law of conservation of energy? Give some examples of each.

Exercise:

Problem:

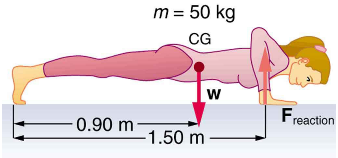
If the efficiency of a coal-fired electrical generating plant is 35%, then what do we mean when we say that energy is a conserved quantity?

Problems & Exercises

Exercise:

Problem: Integrated Concepts

(a) Calculate the force the woman in [link] exerts to do a push-up at constant speed, taking all data to be known to three digits. (b) How much work does she do if her center of mass rises 0.240 m? (c) What is her useful power output if she does 25 push-ups in 1 min? (Should work done lowering her body be included? See the discussion of useful work in Work, Energy, and Power in Humans.



Forces involved in doing push-ups. The woman's weight acts as a force exerted downward on her center of gravity (CG).

Solution:

(a) 294 N

(b) 118 J

(c) 49.0 W

Exercise:

Problem: Integrated Concepts

A 75.0-kg cross-country skier is climbing a 3.0° slope at a constant speed of 2.00 m/s and encounters air resistance of 25.0 N. Find his power output for work done against the gravitational force and air resistance. (b) What average force does he exert backward on the snow to accomplish this? (c) If he continues to exert this force and to experience the same air resistance when he reaches a level area, how long will it take him to reach a velocity of 10.0 m/s?

Exercise:

Problem: **Integrated Concepts**

The 70.0-kg swimmer in [link] starts a race with an initial velocity of 1.25 m/s and exerts an average force of 80.0 N backward with his arms during each 1.80 m long stroke. (a) What is his initial acceleration if water resistance is 45.0 N? (b) What is the subsequent average resistance force from the water during the 5.00 s it takes him to reach his top velocity of 2.50 m/s? (c) Discuss whether water resistance seems to increase linearly with velocity.

Solution:

(a) $0.500\ \mathrm{m/s}^2$

(b) 62.5 N

(c) Assuming the acceleration of the swimmer decreases linearly with time over the 5.00 s interval, the frictional force must therefore be increasing linearly with time, since $f = F - \mathrm{ma}$. If the acceleration decreases linearly with time, the velocity will contain a term dependent on time squared (t^2). Therefore, the water resistance will not depend linearly on the velocity.

Exercise:

Problem: **Integrated Concepts**

A toy gun uses a spring with a force constant of 300 N/m to propel a 10.0-g steel ball. If the spring is compressed 7.00 cm and friction is negligible: (a) How much force is needed to compress the spring? (b) To what maximum height can the ball be shot? (c) At what angles above the horizontal may a child aim to hit a target 3.00 m away at the same height as the gun? (d) What is the gun's maximum range on level ground?

Exercise:

Problem: **Integrated Concepts**

(a) What force must be supplied by an elevator cable to produce an acceleration of $0.800\ \mathrm{m/s}^2$ against a 200-N frictional force, if the mass of the loaded elevator is 1500 kg? (b) How much work is done by the cable in lifting the elevator 20.0 m? (c) What is the final speed of the elevator if it starts from rest? (d) How much work went into thermal energy?

Solution:

(a) 16.1×10^3 N

(b) 3.22×10^5 J

(c) 5.66 m/s

(d) 4.00 kJ

Exercise:

Problem: **Unreasonable Results**

A car advertisement claims that its 900-kg car accelerated from rest to 30.0 m/s and drove 100 km, gaining 3.00 km in altitude, on 1.0 gal of gasoline. The average force of friction including air resistance was 700 N. Assume all values are known to three significant figures. (a) Calculate the car's efficiency. (b) What is unreasonable about the result? (c) Which premise is unreasonable, or which premises are inconsistent?

Exercise:

Problem: **Unreasonable Results**

Body fat is metabolized, supplying 9.30 kcal/g, when dietary intake is less than needed to fuel metabolism. The manufacturers of an exercise bicycle claim that you can lose 0.500 kg of fat per day by vigorously exercising for 2.00 h per day on their machine. (a) How many kcal are supplied by the metabolization of 0.500 kg of fat? (b) Calculate the kcal/min that you would have to utilize to metabolize fat at the rate of 0.500 kg in 2.00 h. (c) What is unreasonable about the results? (d) Which premise is unreasonable, or which premises are inconsistent?

Solution:

(a) $4.65 \times 10^3 \text{ kcal}$

(b) 38.8 kcal/min

(c) This power output is higher than the highest value on [link], which is about 35 kcal/min (corresponding to 2415 watts) for sprinting.

(d) It would be impossible to maintain this power output for 2 hours (imagine sprinting for 2 hours!).

Exercise:

Problem: **Construct Your Own Problem**

Consider a person climbing and descending stairs. Construct a problem in which you calculate the long-term rate at which stairs can be climbed considering the mass of the person, his ability to generate power with his legs, and the height of a single stair step. Also consider why the same person can descend stairs at a faster rate for a nearly unlimited time in spite of the fact that very similar forces are exerted going down as going up. (This points to a fundamentally different process for descending versus climbing stairs.)

Exercise:

Problem: **Construct Your Own Problem**

Consider humans generating electricity by pedaling a device similar to a stationary bicycle. Construct a problem in which you determine the number of people it would take to replace a large electrical generation facility. Among the things to consider are the power output that is reasonable using the legs, rest time, and the need for electricity 24 hours per day. Discuss the practical implications of your results.

Exercise:

Problem: **Integrated Concepts**

A 105-kg basketball player crouches down 0.400 m while waiting to jump. After exerting a force on the floor through this 0.400 m, his feet leave the floor and his center of gravity rises 0.950 m above its normal standing erect position. (a) Using energy considerations, calculate his velocity when he leaves the floor. (b) What average force did he exert on the floor? (Do not neglect the force to support his weight as well as that to accelerate him.) (c) What was his power output during the acceleration phase?

Solution:

(a) 4.32 m/s

(b) $3.47 \times 10^3 \text{ N}$

(c) 8.93 kW

Glossary

renewable forms of energy
: those sources that cannot be used up, such as water, wind, solar, and biomass

fossil fuels
: oil, natural gas, and coal